不確実推論を用いた高信頼度反復復号アルゴリズムの設計と解析

課題番号：12650400

平成12年度～平成14年度科学研究費補助金（基盤研究(C)(2)）研究成果報告書

平成15年5月

研究代表者
松嶋　敏泰（早稲田大学 理工学部 教授）

目次

1. 研究組織

研究代表者：松嶋　敏泰（早稲田大学理工学部教授）
研究分担者：平澤　茂一（早稲田大学理工学部教授）

2. 交付決定額（配分額）

（金額単位：千円）

	直接経費	間接経費	合計
平成 12 年度	1,800	0	1800
平成 13 年度	1,000	0	1,000
平成 14 年度	700	0	700
総計	3,500	0	3,500

3. 研究発表

(1) 学会誌等

著者名	論文標題
後藤正幸，松嶋敏泰，平澤茂一	損失関数を考慮した拡張事後密度の漸近正規性

雑誌名	巻・号	発行年	ページ
電子情報通信学会論文誌	Vol. J83-A, No. 6	2000	639—650

著者名	論文標題
Ken-ichiro Tsukamoto	A Study of the Decision of Control Parameters for Adaptive Automatic-Repeat Request Strategy

雑誌名	巻・号	発行年	ページ
Electronics and Communications in Japan, Part1	Vol.84,No.1	2001	61—70

著者名	論文標題
北原正樹	ウェーブレットパケット基底を用いた信号推定におけるベイズ決定理論の適用に関する一考察

雑誌名	巻・号	発行年	ページ
電子情報通信学会論文誌 A	Vol.J85-A, No.5	2002	584—596

著者名	論文標題
野村亮	メモリ量を低減した近似ベイズ符号化アルゴリズム

雑誌名	巻・号	発行年	ページ
電子情報通信学会論文誌 A	VolJ85-A, No.5	2003	46—59

(2) 口頭発表

著者名	論文標題
Ryo Nomura, Toshiyasu Matsushima, and Shigeichi Hirasawa	On the variance and the probability of length overflow of lossless codes

雑誌名	巻・号	発行年	ページ
Proceedings of IEEE International Symposium on Information Theory		2000	44

著者名	論文標題
Takahiro Yoshida, Toshiyasu Matsushima, and Shigeichi Hirasawa	Achievable rates of random number generators for an arbitrary prescribed distribution from an arbitrary given distribution

雑誌名	巻・号	発行年	ページ
Proceedings of IEEE International Symposium on Information Theory		2000	155

著者名	論文標題
Toshihiro Niinomi, Toshiyasu Matsushima, and Shigeichi Hirasawa	On analysis of noiseless decision feedback scheme using fixed size list decoder for tree codes

雑誌名	巻・号	発行年	ページ
Proceedings of IEEE International Symposium on Information Theory		2000	231

著者名	論文標題
Toshiyasu Matsushima, Tomoko Matsushima and Shigeichi Hirasawa	An interpretation of Turbo decoding from the viewpoint of differential geometry

雑誌名	巻・号	発行年	ページ
Proceedings of International Symposium on Turbo codes & Related Topics,		2000	383—386

著者名	論文標題
鈴木誠，松嶋敏泰，平澤茂一	不確実性を含む演繹推論に関する一考察

雑誌名	巻・号	発行年	ページ
第23回情報理論とその応用シンポジウム予稿集		2000	427—430

著者名	論文標題
野村亮，松嶋敏泰，平澤茂一	ユニバーサル符号における固定長符号の誤り率と可変長符号のオーバーフロー確率について

雑誌名	巻・号	発行年	ページ
第23回情報理論とその応用シンポジウム予稿集		2000	511—513

著者名	論文標題
浮田善文，松嶋敏泰，平澤茂一	フーリエ変換を用いたブール関数の学習に関する一考察

雑誌名	巻・号	発行年	ページ
電子情報通信学会技術研究報告	COMP2000-56	2000	49—55

著者名	論文標題
北原雅樹，野村亮，松嶋敏泰	ウェーブレット・パケットを用いた雑音除去におけるベイズ法の応用に関する一考察

雑誌名	巻・号	発行年	ページ
電子情報通信学会技術研究報告	DSP2000-134	2000	9—16

著者名	論文標題
松嶋敏泰	情報論的学習理論における数理モデル

雑誌名	巻・号	発行年	ページ
人工知能学会誌	Vol. 16, No. 2	2001	252—255

著者名	論文標題
Toshiyasu Matsushima	An Iterative Algorithm for Calculating Posterior Probability and Model Representation

雑誌名	巻・号	発行年	ページ
Proceedings of IEEE International Symposium on Information Theory		2001	—

著者名	論文標題
安田豪毅，野村亮，松嶋敏泰	パラメータが時間変化する情報源とその符号化に関する一考察

雑誌名	巻・号	発行年	ページ
電子情報通信学会技術研究報告	Vol.101, No.177	2001	25—30

著者名	論文標題
吉田隆弘，松嶋敏泰，平澤茂一	多端子モデルに基づく分散協調問題の定式化について

雑誌名	巻・号	発行年	ページ
電子情報通信学会技術研究報告	Vol.101, No.177	2001	37—42

著者名	論文標題
小林学	ブロックターボ符号の生成行列と性能評価

雑誌名	巻・号	発行年	ページ
電子情報通信学会技術研究報告	Vol.101, No.177	2001	1—6

著者名	論文標題
松嶋敏泰	ターボ符号，LDPC 符号の復号アルゴリズム

雑誌名	巻・号	発行年	ページ
ベイジアンネットチュートリアル講演論文集		2001	27—32

著者名	論文標題
桑田修平，吉田隆弘，松嶋敏泰	状態空間モデルを用いた時系列解析に関する一考察ーモンテカルロフィルタにおけるサンプリング方法についてー

雑誌名	巻・号	発行年	ページ
日本経営工学会秋季研究大会予稿集		2001	218—219

著者名	論文標題
吉田隆弘，松嶋敏泰，平澤茂一	多端子情報理論に基づく分散協調問題について

雑誌名	巻・号	発行年	ページ
第24回 情報理論とその応用シンポジウム予稿集		2001	367—370

著者名	論文標題
小林学	ブロックターボ符号に対するインタリーバの構成法と最小距離

雑誌名	巻・号	発行年	ページ
第 24 回 情報理論とその応用シンポジウム予稿		2001	95—98

著者名	論文標題
小林直人	低密度パリティチェック符号の復号アルゴリズムに関する一考察

雑誌名	巻・号	発行年	ページ
電子情報通信学会技術研究報告	Vol.101, No.726	2002	39—44

著者名	論文標題
Toshiyasu Matsushima	An alternate Algorithm for Calculating Generalized Posterior Probability and Decoding

雑誌名	巻・号	発行年	ページ
Proceedings of IEEE International Symposium on Information Theory		2002	338

著者名	論文標題
Toshihiro Niinomi	On the Generalized Viterbi Algorithm using Likelihood Ratio Testing

雑誌名	巻・号	発行年	ページ
Proceedings of IEEE International Symposium on Information Theory		2002	366

著者名	論文標題
Ryo Nomura	On the evaluation of the achievable codelength of Fixed-length codes

雑誌名	巻・号	発行年	ページ
Proceeding of International Sym. on Inf. Theory and Its Application		2002	855—858

著者名	論文標題
松嶋敏泰	一般化事後確率とその計算法

雑誌名	巻・号	発行年	ページ
電子情報通信学会技術報告	IT2002-36	2002	1–8

著者名	論文標題
Toshiyasu Matsushima	Calculation of Generalized Posterior Distribution on Junction Graphs

雑誌名	巻・号	発行年	ページ
第25回情報理論とその応用シンポジウム予稿集		2002	703–706

著者名	論文標題
桑田修平	ベイズ決定理論に基づくロバストなパターン認識に関する一考察

雑誌名	巻・号	発行年	ページ
第25回情報理論とその応用シンポジウム予稿集		2002	283–286

著者名	論文標題
吉田隆弘	多端子情報源符号化に基づいた分散協調問題の定式化

雑誌名	巻・号	発行年	ページ
第25回情報理論とその応用シンポジウム予稿集		2002	351–354

著者名	論文標題
新家稔央	リスト復号に対する判定帰還方式LR+THの誤り指数について

雑誌名	巻・号	発行年	ページ
第25回情報理論とその応用シンポジウム予稿集	IT2002-36	2002	627–630

著者名	論文標題
Ryo Nomura	On the Channel Capacity of Universal Channel Coding

雑誌名	巻・号	発行年	ページ
第25回情報理論とその応用シンポジウム予稿集		2002	635–638

著者名	論文標題
斉藤友彦	誤り訂正符号を用いた直行計画の構成法に関する一考察

雑誌名	巻・号	発行年	ページ
第25回情報理論とその応用シンポジウム予稿集		2002	663–666

著者名	論文標題
石田崇	単語単位で系列を出力する情報源の性質について

雑誌名	巻・号	発行年	ページ
第25回情報理論とその応用シンポジウム予稿集	Vol.J85-A, No.5	2002	695–698

著者名	論文標題				
Toshiyasu Matsushima	Calculation of Generalized Posterior Distribution on Junction Graphs				
雑誌名		巻・号	発行年		ページ
第 25 回情報理論とその応用シンポジウム予稿集			2002		703–706

著者名	論文標題				
須子統太	拡張された階層モデルにおける予測アルゴリズムについて				
雑誌名		巻・号	発行年		ページ
第 25 回情報理論とその応用シンポジウム予稿集			2002		755–758

4. 研究成果

4.1 研究目的

近年，誤り訂正符号の分野で提案されたTurbo符号，Turbo復号法はその誤り訂正能力の高さから21世紀の高信頼度情報通信，蓄積技術の重要な技術として注目を集めている．この符号，復号の優れた性能は幾つかの実験によって検証されているものの，その理論的裏付けはあまり明確にされていなかった．しかし，最近このTurbo復号を含む反復復号アルゴリズムが，知識情報処理の分野で不確実推論アルゴリズムとして用いられている確信度伝搬(BP: Belief Propagation)アルゴリズムの応用例として解釈されることが明らかとなってきた．　このBPアルゴリズムは，不確実な知識をベイジアンネットワークで表し，ある証拠が与えられた場合の各事象の確信度を更新する方法としてとしてPearlにより提案されたアルゴリズムである．事象の確信度は証拠が与えられた元での事後確率に相当し，このBPアルゴリズムは事後確率を計算する効率良いアルゴリズムと位置づけられる．一方，誤り訂正符号において誤り率を最小化する事後確率最大化復号では，受信系列を与えられた元での送信系列の事後確率の計算を効率良く行うことが主要な課題となっており，両者の問題は本質的に同等と考えることができる．Turbo復号はBPアルゴリズムと同等な方法で，事後確率計算を近似的に効率よく行っていると解釈される．しかし，残念ながら複雑な確率構造（グラフ表現ではループのある構造）にBPアルゴリズムを適応した場合の性能は保証されておらず．いまだにTurbo復号法の理論的裏付けは不完全と言わざるを得ない．

そこで本研究では，BPアルゴリズムだけではなく，不確実推論の分野で用いられるアルゴリズムを事後確率計算法として一般化し，有用な符号を含む一般的確率モデルに対して，精度保証がある事後確率計算（復号）アルゴリズムの設計と理論的及び実験的解析を行う．

4.2 一般化確率推論と一般化事後確率

BP（belief propagation）等の通常の確率推論では証拠（evidence）として$X = x$が与えられたもとで，目的とする確率変数Yの事後確率$P(Y|X = x)$を求めている．この場合の証拠は確率変数Xの確定値xが情報として与えられている．本研究では，ある確率変数の分布$P(X = x) = p_x$が情報として与えられた場合の確率推論について考える．また，このような証拠を従来研究ではdistribution-evidenceまたはsoft-evidenceと呼んでいる．

証拠としてdistribution-evidenceが与えれた場合の推論法は従来からいくつか研究されているが，ほとんどの研究がその推論手続きのみを提案しており，推論結果がどのような意味をもつのか，その理論的裏づけを明確にしたものは見当たらない．従来研究のいくつかではJeffry's Ruleと呼ばれる確率計算法を用いて推論を行うことを推奨している．Jeffry's Ruleを用いた計算法は残念なが

ら一つの確率変数に対する distribution-evidence が与えられた場合のみ推論が可能で，複数の確率変数に対して distribution-evidence が与えられた場合には計算が行えない．また，そもそもなぜこの計算法を用いればよいのか，ほとんどの従来研究では定性的概念からしか正当性が説明されていない．

以降，複数の distribution-evidence を扱える一般化確率推論について説明する．これは従来の確率推論や Jeffry's Rule を含む自然な拡張となっている．

確率変数 $X_i, i \in I = 1, ..., n$ と証拠（evidence）と呼ばれる確率変数 $E_j, j \in I_C \subset I$ を定義する．簡単のためこれらの確率変数は離散としておく．一つの証拠 E_j はそれに対応する一つの確率変数 X_j についてのみ情報をもっており，その他の確率変数 $X_i, i \in I - j$ についての情報は一切含まないと考える．これを数式で示すと以下のようになる．

仮定 4.1 確率変数 X_j が与えられたもとで，それに対応する証拠 E_j とそれ以外の確率変数 $X_i, i \in I - j$ は条件付独立になっている．

$$P(X_1, ..., X_n, E_j) = \frac{P(X_1, ..., X_n)P(X_j, E_j)}{P(X_j)}. \tag{1}$$

□

以下で一般化確率推論を定義する．

定義 4.1 証拠 e_j により "X_j の分布は $P^*(X_j)$ である" が与えられたとする．証拠 e_j により与えられる情報は $P^*(X_j) = P(X_j|E_j = e_j)$ とする．この情報が与えられたもとで，事前分布 $P(X_1, ..., X_n)$ から $P(X_1, ..., X_n|E^{I_C} = e^{I_C})$ を求める推論を一般化確率推論と定義する．

□

もし与えられた分布 $P(X_j|E_j = e_j)$ が一点に確率をもつ，すなわち $P(X_j = x_j|E_j = e_j) = 1$ となる場合，証拠 e_j から与えられる情報は "x_j が生起した"，つまり，$X_j = x_j$ と同値である．この場合上記で定義された一般化確率推論は $P(X_1, ..., X_n|X^{I_C} = x^{I_C})$ を求めることと同値になる．ここで $X^{I_C} = x^{I_C}$ は $(X_{j_1} = x_{j_1}, ..., X_{j_{|I_C|}} = x_{j_{|I_C|}})$ を表す．これは通常の確率推論と一致し，一般化確率推論は通常の確率推論を含み，その自然な拡張になっていることがわかる．

定義 4.2 一般化確率推論により求められた分布を，X_j の周辺分布 $P^*(X_j)$ が与えられたもとでの一般化事後分布と定義する．

□

この一般化事後分布がその特殊な場合として通常の事後分布を含むことは明らかであるが，この一般化事後分布が統計的決定理論において，通常の事後分布と全く同様の役割を果たすことに注意されたい．つまり，先に述べた幾つかの決定関数と損失関数に対して，一般化事後確率を最大化する確率変数の値や確率変数の一般化事後確率による期待値などが最適な決定となっている．

次に一般化事後分布と事前分布の重要な関係を示す．

定理 4.1 M_C を X_j の周辺分布が $P^*(X_j) = P(X_j|E_j = e_j), j \in I_C$ を満たす分布 $P(X_1, ..., X_n)$ の集合とする．事前分布 P_{pr} から仮定 4.1 のもとで，定義 4.1 の一般化事後確率推論により求められてた一般化事後確率は以下の式を満たす．

$$P_{po} = \arg \min_{P \in M_C} I(P||P_{pr}). \tag{2}$$

ここで I は Kullback-Leibler(K-L) 情報量を表す．

□

一般化事後確率は，周辺分布の制約のもとで K-L 情報量により測られた距離が事前分布に最も近い分布となっていることを，この定理は示している．

4.3 Itrative Proportional Fitting Procedure （IPFP）

定理 4.1 で示した一般化確率推論の問題は，一種の制約付最適化問題とみなせる．一般に制約付最適化問題において最適会を求めるためには多くの計算量が必要になる．しかし，この周辺分布制約のもとでの K-L 情報量最小化問題については従来からいくつかの研究があり，Itrative Proportional Fitting Procedure（IPFP），または Iterative Scaling Procedure（ISP）と呼ばれる効率的な繰り返しアルゴリズムが提案されている．IPFP はり産データの統計解析において，分割表の周辺和が与えれているもとで，ある種の独立性を仮定したモデルの BAN 推定量や最尤推定量の計算に用いられている．この IPFP は一般化確率推論においても利用可能である．以下で $P(X_1, ..., X_n|E^{I_C} = e^{I_C})$ を求める一般化確率推論の IPFP を適用した手続きを示す．

【Procedure 1A:IPFP】
begin
$P(X_1, ..., X_n) := P_{pr}(X_1, ..., X_n)$;
while $\exists_{j \in I_C} P(X_i) \neq P^*(X_j)$ do
　　begin
　　Pick up X_j from $\{X_j|P(X_j) \neq P^*(X_j), j \in I_C\}$;
　　$P(X_1, ..., X_n) := P(X_1, ..., X_n)P^*(X_j)/P(X_j)$;
　　end
$P_{po}(X_1, ..., X_n) := P(X_1, ..., X_n)$;
end

□

上記の手続きの 6 行目では，周辺分布 $P(X_j)$ が制約条件の周辺分布 $P^*(X_j)$ に合うように毎ステップごとに比例計算により分布の更新を行っている．IPFP はこれを周辺確率が制約条件に収束するまで繰り返す簡潔なアルゴリズムである．

このアルゴリズムの性質を幾何学的に解釈してみる．一つの確率分布を一つの点に対応させた確率分布の空間を考える．周辺分布制約のもとでの事前分布との K-L 情報量を最小化する分布は，事前分布から制約条件を満たす分布の多

様体上へのある種の射影とみなせる．この射影はe-射影またはI-射影と呼ばれている．

補題 4.1 M_C を X_j の周辺分布が $P^*(X_j) = P(X_j|E_j = e_j), j \in I_C$ をみたす分布 $P(X_1, ..., X_n)$ の多様体とする．この多様体はm-平坦となり，一般化事後確率 $P_{po} = \arg\max_{P \in M} I(P||P_{pr})$ は事前分布 P_{pr} から多様体 M_C 上へのe-射影で与えられる．

また，任意の分布 $p \in M_C$ は以下を満たす．

$$I(p||P_{pr}) = I(p||P_{po}) + I(P_{po}||P_{pr}). \tag{3}$$

□

複数の周辺制約を満たす多様体へのe-射影は，ここの周辺制約へのe-射影の繰り返しにより求めることができる．

補題 4.2 複数の周辺制約 $C = \bigcap_{j=1}^{m} C_j \neq \phi$ を構成する一つの周辺制約 C_j を満たす分布の多様体を M_j で表す．分布 P_t を分布 P_{t-1} から多様体 $M_t, t = 1, 2, ...$ へのe-射影とする．ここで，$P_0 = P_{pr}$ であり，$M_t = M_j, (\mathrm{mod}\ m)$ とする．以上により再帰的に定義された $P_1, P_2, ...$ の系列は P_{pr} から周辺制約 C を満たす多様体 M_C 上へのe-射影に収束する．

□

事前分布 P_{pr} から複数の周辺制約を満たす多様体上へのe-射影は以下のような周辺パラメータを用いて表すことが可能である．

補題 4.3 事前分布 P_{pr} から周辺制約 $P(x_j) = P^*(x_j), j \in I_C$ を満たす多様体上へのe-射影である一般化事後分布 P_{po} は以下のように表すことが出来る．

$$P_{po}(x_1, ..., x_n) = P_{pr}(x_1, ..., x_n) \prod_{j \in I_C} \beta(x_j), \tag{4}$$

ここで

$$P^*(x_j) = \sum_{\{x_i | i \neq j\}} P_{pr}(x_1, ..., x_n) \prod_{j \in I_C} \beta(x_j). \tag{5}$$

□

系 4.1 $P(X_1, ..., X_n)$ から周辺制約 $P^*(X_j)$ を満たす多様体上へのe-射影

$$P^*(X_1, ..., X_n)$$

は以下のように計算される．

$$P^*(x_1, ..., x_n) = P(x_1, ..., x_n) \frac{P^*(x_j)}{P(x_j)}. \tag{6}$$

□

Procedure 1A つまりIPFPの正当性は補題4.2と系4.1より以下のように示される．

補題 4.4 Procedure 1A は停止し，求められた分布は $P(X_1,...,X_n|E^{I_C}=e^{I_C})$ へ収束する．

□

4.4 周辺パラメータを用いたアルゴリズムと並列アルゴリズム

Procedure 1A においては，計算過程の分布を結合分布 $P(X_1,...,X_n)$ で表現し保持していたが，補題 4.3 で示したように，目的とする事後分布は事前分布 P_{pr} と周辺パラメータ $\beta(X_j), j \in I_C$ を用いても表現可能である．

Procedure 1A を事前分布 P_{pr} と周辺パラメータ $\beta(X_j)$ を用いた表現で書き直すことで，以下のような Procedure 1B が提案されている．

【Procedure 1B】
begin
$\beta(X_j) := 0, j \in I_C$
while $\exists_{j \in I_C} P(X_i) \neq P^*(X_j)$ do
　　begin
　　Pick up X_j from $\{X_j | P(X_j) \neq P^*(X_j), j \in I_C\}$;
　　$\gamma(X_j) := \sum_{i \neq j} P_{pr}(X_1,...,X_n) \prod_{j \in I_C} \beta(X_j)$;
　　$\beta(X_j) := P^*(X_j)/\gamma(X_j)$;
　　end
$P_{po}(X_1,...,X_n) := P_{pr}(X_1,...,X_n) \prod_{j \in I_C} \beta(X_j)$;
end

□

このアルゴリズムを Procedure 1A と比較すると，収束するまでの反復回数は同じであるが，メモリ量は少なくて済む．また，後で述べる分解可能モデルに対して Procedure 1B を利用した確率伝播アルゴリズムは Procedure 1A を利用した場合に比べ計算量が少ないという特長がある．

Procedure 1A と Procedure 1B は逐次アルゴリズムであるが，以下のような並列アルゴリズムも提案されている．

【Procedure 2B】
begin
$\beta(X_j) := 0, j \in I_C$
while $\exists_{j \in I_C} P(X_i) \neq P^*(X_j)$ do
　　begin
　　Pick up X_j from $\{X_j | P(X_j) \neq P^*(X_j), j \in I_C\}$;
　　$\gamma(X_j) := \sum_{i \neq j} P_{pr}(X_1,...,X_n) \prod_{j \in I_C} \beta(X_j), j \in I_C$;
　　$\beta(X_j) := P^*(X_j)/\gamma(X_j)$;
　　end
$P_{po}(X_1,...,X_n) := P_{pr}(X_1,...,X_n) \prod_{j \in I_C} \beta(X_j), j \in I_C$;

end □

Procedure 1A と Procedure 1B は一回の繰り返しにおいて，一つの周辺分布制約についてe-射影を行うことで分布の更新を行っているが，この並列アルゴリズムでは全ての周辺分布制約に対して同時に更新を行っていることになる．Procedure 2B が行っている計算を幾何学的に解釈してみよう．1時点前のステップで計算された分布から，各周辺制約を満たす多様体へのe-射影を考える．この分布から各e-射影へのベクトルを用い，それらの合成ベクトルを作る．この合成ベクトルにより移される分布がこのステップで計算された分布となっていることが分る．この解釈を厳密化することで以下の定理が証明される．

定理 4.2 Procedure 2B は停止し，Procedure 1A および Procedure 1B と同様の分布に収束する．

□

4.5 Junction graph の拡張

確率モデルが次式のようにクリークの積で記述される場合，従来の Junction graph(JG) を拡張した extended junction graph （EJG） が提案されている．[7)7)]

$$P(x_1, ..., x_n) = \alpha q(N_1)q(N_2)\cdots q(N_{n_N}), \tag{7}$$

ここで，$N_l = (X_{i_1(l)}, ..., X_{i_{n(l)}(l)})$.

EJG は JG と同様に交差節 (intersection node) とクリーク節 (clique node) の2種類の節で記述されるが，交差節の結合法が少々異なっている．従来の JG では交差節は必ず2つのクリーク節と枝で結ばれているが，EJG の場合は2つ以上のクリーク節と結合されても構わない．JG の場合，全ての交差節を削除しても残されたグラフは通常のグラフであるが，EJG の場合，残されたグラフが hypre graph となる場合もありえる．

EJG や JG は BN や factor graph を含む，より広い範囲の確率モデルを表現できるグラフと考えられる．

4.6 EGJ 上の一般化事後確率の効率的計算法

一般の確率モデルにおける一般化事後確率の計算は確率変数の数に対して指数オーダの計算量が一回の繰り返しにおいて必要であり，通常の事後確率計算以上に膨大な計算量が必要である．しかし，通常の事後確率の計算と同様に，確率モデルが式 (7) のようにクリークの積で記述される場合，効率的アルゴリズムが提案されている．従来の周辺事後確率計算と同様にグラフで確率モデルを表現しその上でメッセージを伝播させる方法が有効である．

ループのない EJG 上での周辺一般化事後確率の計算は，Procedure 1B を各クリーク節ごとに行い，その結果を隣接するノードに伝播させることで基本的には計算が行える．これは逐次的確率伝播アルゴリズムと位置付けられる．

他方，Procedure 2B を応用することで，一般化事後確率を計算する並列確率

伝播アルゴリズムも構成することができる．この並列アルゴリズムはループのあるEJGにもそのまま適用することが可能となり，正確な一般化事後確率計算のみならず近似計算にも利用可能である．

ここでは並列伝播アルゴリズムを示す．このアルゴリズムでは交差節からクリーク節へのメッセージと，その逆にクリーク節から交差節へのメッセージが交互に交換される．

【Algorithm 1B】

各交差節 D_m からそれに隣接するクリーク節 N_i へのメッセージの伝播は以下となる．$D_m = \{X_j\}$ かつ $j \in I_C$ のとき，D_m を制約節と呼ぶ．また，$S^N(D_m)$ は D_m に隣接するクリーク節の集合とする．

$$q_{t+1}^{N_l}(D_m) := \begin{cases} P^*(D_m)/\gamma_t^{N_l}(D_m), & D_m\text{は制約節} \\ \prod_{\{N_k \in S^N(D_m)|k \neq l\}} \gamma_t^{N_k} D(_m), & \text{その他} \end{cases} . \tag{8}$$

各クリーク節 N_l からそれに隣接する交差節 D_m へのメッセージ伝播は以下となる．$S^D(N_l)$ は N_l に隣接する交差節の集合とする．

$$\gamma_t^{N_k} D(_m) := \sum_{x \notin D_m} q(N_l) \prod_{\{D_h \in S^D(N_l)|h \neq m\}} q_t^{N_l}(D_h). \tag{9}$$

それぞれのクリーク節の周辺一般化事後確率は以下のように計算される．

$$P_{t+1}(N_l) = q(N_l) \prod_{D_h \in S^D(N_l)} q_t^{N_l}(D_h). \tag{10}$$

□

定理 4.3 もしEJGがループを含まなければ，Algorithm 1Bは停止し，正しい一般が事後確率を計算する．

□

一般化事後確率は通常の事後確率も含んでいるため，このアルゴリズムにより通常の事後確率も計算可能である．その意味で，このアルゴリズムはBPやHUGINアルゴリズムの一般化になっている．さらにEJGがBNやfactor graphを含むことから，表現できる確率モデルのクラスの意味でも拡張されている．

この並列伝播アルゴリズムは確率推論の分野における distribution-evidence を含む推論への応用はもちろんのこと，復号問題など通常の事後確率計算を行っている分野への応用も可能である．

4.7 LDPC 符号への適用

EJGやJGはBNやfactor graphを含む，より広い範囲の確率モデルを表現できるグラフと考えれらる．例えばLDPC符号はfactor graphで記述されることが一般的であるが，多くのループを含んだグラフとなる．特に長さ4のループがfactor fraphに含まれるとsum-productアルゴリズムは極端に性能が劣化することが知られており，通常は長さ4のループを含まないようにLDPC符号は構成される．EJGやJGを用いた場合，factor graphによる表現では長さ4

のループを含む確率モデルでもループなしで表現可能であり，Algorithm 1B 等を用いることで正確な事後確率が計算可能である．よってEJGやJGを用いて長さ4のループを含むLDPC符号を表現することにより，その部分による確率伝播アルゴリズムの精度劣化の影響を軽減することが可能であると予測される．これは既にいくつかの実験によって確かめられており，長さ4のループを含むLDPC符号に対して，factor graph で表現し sum-product アルゴリズムを用いたものに比べてもEJGで表現した場合のほうが誤り率が低いことが報告されている

このことはよりランダムに近い構成法を持つ符号のほうが符号の持つ特性がよくなることを具体的に示したといえる．

4.8 畳み込み符号への適用

畳み込み符号，tail-biting 畳み込み符号に Algorithm 1B を適用した場合についてもいくつかの実験によってその性能が確かめられている．その結果，BCJR 復号と同じ誤り率をより短い時間で得られること，復号に必要な時間は畳み込み符号の拘束長に比例することなどが分っている．

この実験を踏まえ，畳み込み符号の場合の性能の理論値を数値計算によって示した．また Algorithm 1B における sum 演算を max 演算に変換することによって Viterbi アルゴリズムの並列計算版を構成可能である．これを tail-biting 畳み込み符号に適用してブロック誤り率を調べる実験を行った．結果 tail-biting Viterbi アルゴリズム，tail-biting BCJR アルゴリズムを用いるよりも計算時間は大幅に減少し，ブロック誤り率も向上することを示した．

4.9 ブロック符号への適用

より密な符号に対しての性能評価も行った．従来から様々な所で研究されてきたBCH符号やRS符号といった一般的なブロック符号を，factor graph で表現すると長さ4のループを無数に含む．そのため sum-product アルゴリズム等の繰り返し復号アルゴリズムを適用してもあまり良い性能は得られなかった．しかし，この場合もLDPC符号と同様にEJGやJGを用いて復号することにより長さ4のループによる復号性能の劣化を抑えられることが期待される．

そこで，Algorithm 1B 等を適用するため，一般的なブロック符号を復号する場合のEJGの構築方法を提案し，これを用いて一般的なブロック符号 (BCH 符号の双対符号) を反復復号によって復号する研究を行った．この場合も Sum-Product アルゴリズムで復号するときよりも復号誤り率が改善されることが分かった．

5. 関 連 文 献

損失関数を考慮した拡張事後密度の漸近正規性

後藤　正幸[†a)]　松嶋　敏泰[††]　平澤　茂一[††]

Asymptotic Normality of Extended Posterior Density Functions with Loss Functions

Masayuki GOTOH[†a)], Toshiyasu MATSUSHIMA[††], and Shigeichi HIRASAWA[††]

あらまし　漸近正規性は統計的推測における本質的な性質である．例えば，J. Rissanen は最ゆう推定量の漸近正規性のもとで，ユニバーサル符号化，ユニバーサル予測，ユニバーサルモデル化に関する重要な性質を導いている．本論文では，損失関数を考慮した場合のパラメータの拡張事後密度を定式化する．従来，ベイズルールに基づく事後確率密度や拡張確率的コンプレキシティ(ESC)のための拡張事後密度が示されているが，本論文で定義する拡張事後密度は，損失に対し単調減少，事前確率に対し単調増加のみを仮定しただけの密度関数である．更に，漸近正規性のために必要な密度関数のもつべき条件を示す．これにより漸近正規性という観点から，損失関数を考慮した場合の拡張事後密度として望ましい形が再認識される．本論文で議論する拡張事後密度の漸近正規性は，確率測度の法則収束の意味とは異なるが，適当な条件のもとでベイズ事後確率密度が漸近正規性を満たすことが知られているように，本質的な性質といえる．本論文ではこの漸近正規性がいくつかの重要な結果を導くことを示す．

キーワード　拡張事後密度，漸近正規性，拡張確率的コンプレキシティ

1. まえがき

統計的推測において，適当な次元のパラメータをもつパラメトリックな確率モデルを仮定することがしばしば行われる．そして実際に得られたデータ系列から，そのパラメータに関して推定を行ったり，次に得られるデータの予測を行うために用いられる．実際に有用な方法の一つは，データ系列から何らかの形でパラメータを推定し，それを目的に応じて利用することであろう．

確率モデルのパラメータ推定としては，点推定，区間推定など多くの定式化があるが，本論文では事前分布が仮定できる場合を考える．パラメータ空間上に事前分布が仮定できれば，ベイズ規則により事後分布を得ることができ，ベイズ統計理論に基づく定式化が可能となる [1]．これは理論的にも非常に明快な解答を与える．一方で，データ系列からパラメータの推定量として，分布の形で推定量を得る方法が考えられる [22]．その一つはベイズ規則に基づく事後確率密度である．本論文では，統計的推測において損失関数を考慮した場合について，密度関数の形式でのパラメータ推定を定式化する．ここで得られる密度を拡張事後密度と呼ぶが(注1)，本論文で定義する拡張事後密度は，損失に対し単調減少，事前確率に対し単調増加のみを仮定しただけの密度関数である．更に，この拡張事後密度が漸近的に正規分布に収束するための関数の満たすべき条件を示す．

漸近正規性は統計的推測において本質的であり，J. Rissanen は最ゆう推定量の漸近正規性のもとで，ユニバーサル符号化，ユニバーサル予測，ユニバーサルモデル化の問題に対する重要な性質を導いている [16], [18]．最ゆう推定量の場合は，真の確率測度に基づく法則収束の意味での漸近正規性が議論される．一方，ベイズ規則に基づいて得られる事後確率密度が，適当な条件の

† 東京大学大学院工学系研究科，東京都
School of Engineering, The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo, 133–8656 Japan
†† 早稲田大学理工学部経営システム工学科，東京都
School of Science and Engineering, Waseda University, 3–4–1 Ohkubo, Shinjuku-ku, Tokyo, 169–8555 Japan
a) E-mail: goto@naoe.t.u-tokyo.ac.jp

(注1)：本論文では，ベイズの定理に基づいて計算される事後確率密度に対してのみ確率という言葉を付けて用いる．本論文で定義する密度関数は，損失関数に基づいて計算されるもので厳密な意味での事後確率分布を表していないので，単に事後密度と呼んでいる．

もとで漸近正規性を満たすことも知られている (A.M. Walker [21], C.C. Heyde and I.M. Johnston [11] など). 事後確率密度は個々のデータ系列に対応して得られるので, 確率収束, 概収束, 一様収束などの意味での漸近正規性が議論の対象となるが [13], この性質からもベイズ統計に基づく方法の漸近的評価が可能であるなど [3], [9], 重要な性質を導くことができる [1], [5]. 拡張事後密度の漸近正規性はこのような議論の一般化となっている. その結果, 拡張事後密度が損失に対し指数減少以上であるときに漸近正規性が成り立つことを示す. 従来から議論されている拡張事後密度は損失に対し指数減少の形のものであり, 確率論的にはこれが自然である. しかし, より広いクラスの拡張事後密度の性質を議論することによって, 漸近正規性という観点から論理的に損失関数を考慮した場合の拡張事後密度として望ましい形が再認識される.

本論文では補足として, 前半の結果から K. Yamanishi により提案された拡張確率的コンプレキシティ [23], [24] の漸近評価が可能となり, その意味が漸近正規性の面から再認識されるということを示す. ESC に関する概収束と平均収束の結果は [9] の議論と同じであるが, 前半で定義する拡張事後密度が漸近正規性を満たすための条件の境界に位置するものが ESC であるという観点から, ESC の正当性について考察する. 更に, データ系列のある部分集合について成り立つ一様収束の意味での漸近式についても述べる.

2. 準　　備

2.1 1本の密度関数列の漸近正規性

まず最初に 1 本の Θ^k 上の密度関数列 $f_n(\theta^k)$, $n = 1, 2, 3, \cdots$ を考える. ただし $\theta^k \in \Theta^k$ かつ $\int_{\theta^k} f_n(\theta^k) d\theta^k = 1$, $n = 1, 2, 3, \cdots$ とする. $\Theta^k \subset \mathcal{R}^k$ とし, $\mathcal{R}^k$ は k-次元ユークリッド空間とする.

$\tilde{\theta}^k$ は $K_n(\theta^k) = \log f_n(\theta^k)$ のある極大値とし,

$$K_n'(\tilde{\theta}^k) = \mathbf{0}, \tag{1}$$

を満たし, かつ

$$\Sigma_n = -\left(K_n''(\tilde{\theta}^k)\right)^{-1} \tag{2}$$

は正定値行列であるとする. ただし $K_n'(\tilde{\theta}^k)$ と $K_n''(\tilde{\theta}^k)$ は

$$K_n'(\tilde{\theta}^k) = \left.\frac{\partial K_n(\theta^k)}{\partial \theta^k}\right|_{\theta^k = \tilde{\theta}^k} \tag{3}$$

$$K_n''(\tilde{\theta}^k) = \left.\frac{\partial^2 K_n(\theta^k)}{\partial \theta^k (\partial \theta^k)^T}\right|_{\theta^k = \tilde{\theta}^k} \tag{4}$$

であり, T は行列またはベクトルの転値である.

[補題 1] [1], [5] 球 $B_\delta(\tilde{\theta}^k) = \left\{\theta^k \in \Theta^k \middle| \|\theta^k - \tilde{\theta}^k\| < \delta\right\}$ を定義すると, 次の三つの条件が密度関数 $f_n(\theta^k)$ の漸近正規性の必要十分条件である.

(c.1) *"Steepness"* $\lim_{n\to\infty} \bar{\sigma}_n^2 \to 0$, ただし $\bar{\sigma}_n^2$ は Σ_n の最大固有値である.

(c.2) *"Smoothness"* 任意の $\epsilon > 0$ に対し, 自然数 N と $\delta > 0$ が存在し, $\forall n > N$ と $\forall \theta^k \in B_\delta(\tilde{\theta}^k)$ に対し $K_n''(\theta^k)$ が存在して, かつ

$$I - A(\epsilon) \leqq K_n''(\theta^k)\left\{K_n''(\tilde{\theta}^k)\right\}^{-1} \leqq I + A(\epsilon) \tag{5}$$

が $\forall \theta^k \in B_\delta(\tilde{\theta}^k)$ に対し一様に成り立つような $A(\epsilon)$ が存在する. ただし, I は $k \times k$ 単位行列, $A(\epsilon)$ は $\epsilon \to 0$ のときのその最大固有値が 0 に収束するような $k \times k$ 次元の非負定値行列とする.

(c.3) *"Concentration"* 任意の $\delta > 0$ に対し

$$\int_{\theta^k \in B_\delta(\tilde{\theta}^k)} f_n(\theta^k) d\theta^k \to 1 \tag{6}$$

条件 (c.1), (c.2), (c.3) は

$$f_n(\tilde{\theta}^k)\,(\det \Sigma_n)^{1/2} \to (2\pi)^{-k/2} \tag{7}$$

を意味する. 更に, 条件 (c.1), (c.2), (c.3) のもとで, $\phi_n^k = \Sigma_n^{-1/2}(\theta^k - \tilde{\theta}^k)$ の密度関数, $f_{\phi,n}(\phi_n^k)$, は

$$\int_R f_{\phi,n}(\phi_n^k) d\phi_n^k \to \int_R (2\pi)^{-k/2} \exp\left\{-\frac{1}{2}(\phi_n^k)^T \phi_n^k\right\} d\phi_n^k \tag{8}$$

を満たす. ただし, R は任意の直方体であり, $f_{\phi,n}(\phi_n^k)$ は

$$f_{\phi,n}(\phi_n^k) = (\det \Sigma_n)^{1/2} f_n(\theta^k) \tag{9}$$

で与えられる. □

この補題は 1 本の密度関数列に対する結果であり, 一般に式 (8) の性質を漸近正規性という. ただし, 以下では紙面の都合のため, 後の議論に必要な式 (7) の形の収束のみを結果として示す. 式 (7) が成り立てば, ほぼ同様の手順によって式 (8) も得られる [1]. また, 本論文ではデータ系列から計算されるパラメータ空間上の密度関数を議論するが, これは 1 本の密度関数列を考えているわけではない.

2.2 データ系列により与えられる事後密度の漸近正規性

補題 1 は統計的推測に利用できる形に拡張できる．すなわち，長さ n のデータ系列 x^n に依存する密度関数列 $f_{x^n}(\theta^k)$ を考える．ここで，$x^n = x_1 x_2 \cdots x_n$ は確率変数 $X^n = X_1 X_2 \cdots X_n$ の実現値であり，真の確率分布 $p^*(x^n)$ に従う長さ n の i.i.d. 系列であるとする．$\mathcal{X}$ をデータ空間，$x_i \in \mathcal{X}$ とする．また，$x^n \in \mathcal{X}^n$，すなわち $\mathcal{X}^n$ はデータ系列 x^n の取り得る空間，すなわち確率変数 X^n の値域とする．補題 1 より，以下の補題が明らかに成り立つ．

[補題 2] $\tilde{\theta}^k = \tilde{\theta}^k(x^n)$ を $K_{x^n}(\theta^k) = \log f_{x^n}(\theta^k)$ の極大値とし，

$$K'_{x^n}\left(\tilde{\theta}^k\right) = \mathbf{0} \tag{10}$$

かつ

$$\Sigma_{x^n} = -\left(K''_{x^n}(\tilde{\theta}^k)\right)^{-1} \tag{11}$$

は正定値であるとする．

(c.1') *"Steepness"* $\lim_{n\to\infty} \bar{\sigma}^2_{x^n} \to 0$ *a.s.*, ただし $\bar{\sigma}^2_{x^n}$ は Σ_{x^n} の最大固有値であり，*a.s.* は概収束を意味する．

(c.2') *"Smoothness"* 任意の $\epsilon > 0$ に対し，$\delta > 0$ が存在して，任意の $\theta^k \in B_\delta(\tilde{\theta}^k)$ に対して $L''_{x^n}(\theta^k)$ が存在して $n \to \infty$ のとき

$$I - A(\epsilon) \leqq K''_{x^n}(\theta^k)\left\{K''_{x^n}(\tilde{\theta}^k)\right\}^{-1} \leqq I + A(\epsilon), \quad a.s. \tag{12}$$

が成り立つ(注2)．

(c.3') *"Concentration"* 任意の $\delta > 0$ に対し

$$\int_{\theta^k \in B_\delta(\tilde{\theta}^k)} f_{x^n}(\theta^k) d\theta^k \to 1, \quad a.s. \tag{13}$$

条件 (c.1)，(c.2)，(c.3) は

$$f_{x^n}(\tilde{\theta}^k)\left(\det \Sigma_{x^n}\right)^{1/2} \to (2\pi)^{-k/2}, \quad a.s. \tag{14}$$

を意味する．

条件 (c.1')，(c.2')，(c.3') の収束を $p^*(\cdot)$ に対する確率収束で置き換えると，

$$f_{x^n}(\tilde{\theta}^k)\left(\det \Sigma_{x^n}\right)^{1/2} \to (2\pi)^{-k/2}, \quad in\ prob. \tag{15}$$

が成り立つ．ただし，*in prob.* は確率収束を意味する．

条件 (c.1')，(c.2')，(c.3') の収束を $p^*(\cdot)$ に対する平均収束で置き換えると，

$$E^*\left[f_{x^n}(\tilde{\theta}^k)\left(\det \Sigma_{x^n}\right)^{1/2}\right] \to (2\pi)^{-k/2} \tag{16}$$

が成り立つ．ただし，$E^*[\cdot]$ は $p^*(\cdot)$ に基づく期待値を表す．

データ系列のある部分集合 $\overline{\mathcal{X}}^n \subset \mathcal{X}^n$ を考えたとき，条件 (c.1')，(c.2')，(c.3') が，データ系列 $x^n \in \overline{\mathcal{X}}^n$ に対して一様に成り立つ収束で置き換えられるものとすると，$x^n \in \overline{\mathcal{X}}^n$ に対して一様に

$$f_{x^n}(\tilde{\theta}^k)\left(\det \Sigma_{x^n}\right)^{1/2} \to (2\pi)^{-k/2} \tag{17}$$

が成り立つ． □

3. 主 結 果

3.1 損失関数に基づく拡張事後密度

補題 2 では密度関数 $f_{x^n}(\theta^k)$ はデータ系列 x^n に依存するものとしたが，実際にどのように計算されるかについては言及しなかった．ここでは，データ系列からパラメータ空間上の密度関数の形で推定量を得る問題を考え [22]，なるべく一般的な定式化を試みる．

パラメータ θ^k で特長づけられる確率モデルを $p_{\theta^k}(\cdot)$ とし，確率モデル族を $\{p_{\theta^k}(\cdot)|\theta^k \in \Theta^k\}$ と記述する．$\pi(\theta^k)$ を Θ^k 上の事前確率密度とする．事前確率密度 $\pi(\theta^k)$ のもとで，データ系列 x^n を得たときに推定量である $f_{x^n}(\theta)$ を計算する方法を考えよう．すなわち，分布の形でパラメータの推定量を得る方法について考える．

例えば，ベイズ規則から得られる事後確率密度 $f_B(\theta^k|x^n)$ は

$$f_B(\theta^k|x^n) = \frac{p_{\theta^k}(x^n)\pi(\theta^k)}{\displaystyle\int_{\theta^k} p_{\theta^k}(x^n)\pi(\theta^k)d\theta^k} \tag{18}$$

と与えられる．以下では，これをベイズ事後密度と呼ぶ．

次に，これを任意の損失関数を考慮できる場合に拡張しよう．

$\mathcal{D}$ はユークリッド空間の部分集合，あるいは X 上の確率分布の集合を表すものとする．$\mathcal{D}$ を決定空間と呼び，その要素を d とし決定と呼ぶ．統計的推測のためにパラメタライズされた仮説空間 $\mathcal{H} = \{d_{\theta^k}|\theta^k \in \Theta^k\}$ を準備する．$\mathcal{H} \subseteq \mathcal{D}$ であり，要素である d_{θ^k} を一つの

(注2)：本論文では，ある事象 A_n に対し $P^*(\cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k) = 1$ であるとき，"$n \to \infty$ のとき A_n, *a.s.* である" と記述する．確率収束や一様収束に関しても同様．

仮説（あるいは決定）といい，$\mathcal{H}$ は実数の集合，あるいは確率分布の集合である．Θ^k は k 次元パラメータ空間であり，k 次元ユークリッド空間 $\mathcal{R}^k$ の部分集合とする．

例えば確率分布 $p(x)$ を決定する問題の場合，$\mathcal{D} = \left\{p(x) \left| \int_x dp(x) = 1 \right.\right\}$ となる．特に $\mathcal{X}$ が離散集合の場合は $\mathcal{D} = \left\{p(x) \left| \sum_x p(x) = 1 \right.\right\}$ である．これに対して仮定される仮説空間が $\mathcal{H} = \{p_{\theta^k}(x) | \theta^k \in \Theta^k\}$ となる．

X の値を予測する問題なら，$\mathcal{D} = \mathcal{X}$ とすればよい．[23] では，二つの確率変数間の関係を議論する問題，すなわち X を得て Y の値や分布を予測するケースを扱っているが，本論文ではより単純に X のみを考えている．もしこのような場合を考えるならば，$\mathcal{X} \to \mathcal{Y}$ と写像する関数の集合を仮説空間とし $\mathcal{D} = \{f(Y|X)\}$，$\mathcal{H} = \{f_{\theta^k}(Y|X)\}$ などとすればよい．

$L(d : x) : \mathcal{D} \times \mathcal{X} \to [0, \infty)$ を決定 d と $x \in \mathcal{X}$ 間の損失関数とする．例えば，決定空間 $\mathcal{D}$ が確率モデルのクラス $\{p(X)\}$ であるとき，すなわち $\mathcal{H} = \{p_{\theta^k}(X) : \theta^k \in \Theta^k\}$ であるときには，損失関数は $L(p_{\theta^k}(\cdot) : X)$ のように与えられる．例えば X が離散確率変数のとき，α-損失関数は

$$L(p_{\theta^k}(\cdot) : X) = \left(1 - p_{\theta^k}(X)\right)^{\alpha} \tag{19}$$

で与えられ(注3)，対数損失関数は

$$L(p_{\theta^k}(\cdot) : X) = -\log p_{\theta^k}(X) \tag{20}$$

で与えられる．その他の決定空間や損失関数については [23] を参照．

本論文では，次の形の拡張事後密度 $f_{x^n}(\theta^k)$ を考える．

$$f_{x^n}(\theta^k) = \frac{g\left(\sum_{t=1}^{n} L(d_{\theta^k} : x_t), \pi(\theta^k)\right)}{\int_{\theta^k} g\left(\sum_{t=1}^{n} L(d_{\theta^k} : x_t), \pi(\theta^k)\right) d\theta^k} \tag{21}$$

ここで，関数 $g(y, z)$ は $y \in [0, +\infty)$ に関して単調減少，$z \in [0, +\infty)$ に関して単調増加な非負関数とする．その意味するところは，事前確率密度が高く，損失が小さいパラメータには高い事後密度が割り当てられるということである．ここで，$g'(y, z) = \frac{\partial g(y,z)}{\partial y}$ と記述することにすると，$g(y, z)$ は非負であるから下に有界であるので，$y \to \infty$ のとき

$$g'(y, z) = O(g(y, z)) \tag{22}$$

であることに注意しておく(注4)．なぜなら，平均値の定理より $\forall y_1, \forall y_2 \in [0, +\infty)$ に対して

$$g(y_1, z) = g(y_2, z) + g'(y_3, z)(y_1 - y_2)$$

かつ $y_1 \geqq y_3 \geqq y_2$ となる $y_3 \in [0, +\infty)$ が存在するので，$y_2 \to \infty$ を考えたとき，$\left|\frac{g'(y,z)}{g(y,z)}\right| \to \infty$ とすると $g(y, z)$ は有界となり得なくなってしまうからである．

もし $g(y, z) = \frac{\lambda}{y} + z$ であれば

$$f_{x^n}(\theta^k) = \frac{\frac{\lambda}{\sum_{t=1}^{n} L(d_{\theta^k} : x_t)} + \pi(\theta^k)}{\int_{\theta^k} \left(\frac{\lambda}{\sum_{t=1}^{n} L(d_{\theta^k} : x_t)} + \pi(\theta^k)\right) d\theta^k} \tag{23}$$

となる．類似の形として

$$f_{x^n}(\theta^k) = \frac{\sum_{t=1}^{n} \frac{\lambda}{L(d_{\theta^k} : x_t)} + \pi(\theta^k)}{\int_{\theta^k} \left(\sum_{t=1}^{n} \frac{\lambda}{L(d_{\theta^k} : x_t)} + \pi(\theta^k)\right) d\theta^k} \tag{24}$$

が考えられる．

もし $g(y, z) = z \exp\{-\lambda y\}$ であれば

$$f_{x^n}(\theta^k) = \frac{\exp\left(-\lambda \sum_{t=1}^{n} L(d_{\theta^k} : x_t)\right) \pi(\theta^k)}{\int_{\theta^k} \exp\left(-\lambda \sum_{t=1}^{n} L(d_{\theta^k} : x_t)\right) \pi(\theta^k) d\theta^k} \tag{25}$$

となる．これは K. Yamanishi [23] により提案された ESC に対して自然に導かれる事後密度である．

(注3)：式 (19) は X が離散確率変数の場合の α-損失関数である．例えば $\mathcal{D} = \mathcal{X}$ の場合には，$|X - d|^{\alpha}$ と定義される．

(注4)：本論文を通じて次の記法を用いる．$g(y)$, $h(y)$ を y の関数とし，$g(y) = O(h(y))$ は，$\forall y \geqq y_0$ に対し $|g(y)| \leqq c|h(y)|$ となるような正定数 c, y_0 が存在することを意味する．また，$g(y) = o(h(y))$ は $\lim_{y \to \infty} \frac{g(y)}{h(y)} = 0$，$g(y) = \Omega(h(y))$ は $g(y) \neq o(h(y))$ であることを意味する．

ある損失関数 $L(d_{\theta^k}:x)$ に対し，$L_1(d_{\theta^k}:x)=\exp\{L(d_{\theta^k}:x)\}$ と定義すると，式 (25) は

$$f_{x^n}(\theta^k)=\frac{\prod_{t=1}^{n}(L_1(d_{\theta^k}:x_t))^{-\lambda}\pi(\theta^k)}{\int_{\theta^k}\prod_{t=1}^{n}(L_1(d_{\theta^k}:x_t))^{-\lambda}\pi(\theta^k)d\theta^k} \tag{26}$$

と書き換えられる．したがって，式 (23) と式 (25) はそれぞれ，損失の和と積に比例する形の密度関数を与えている．

もし $g(y,z)=z^{\lambda}\exp\{-\lambda y\}$ とすれば，式 (26) と類似した形として，

$$f_{x^n}(\theta^k)=\frac{\prod_{t=1}^{n}\left\{L_1(d_{\theta^k}:x_t)\pi(\theta^k)\right\}^{\lambda}}{\int_{\theta^k}\prod_{t=1}^{n}\left\{L_1(d_{\theta^k}:x_t)\pi(\theta^k)\right\}^{\lambda}d\theta^k} \tag{27}$$

が導かれる．$L_1(d_{\theta^k}:x_t)=p_{\theta^k}(x_t)$ であるとき，ボルツマン分布

$$f_{x^n}(\theta^k)=\frac{\left\{p_{\theta^k}(x^n)\pi(\theta^k)\right\}^{\lambda}}{\int_{\theta^k}\left\{p_{\theta^k}(x^n)\pi(\theta^k)\right\}^{\lambda}d\theta^k} \tag{28}$$

が得られる [22]．もし $\lambda=1$ であれば，ベイズ事後密度となる．$\lambda=\infty$ と考えれば，これは事後確率最大推定量と等価になる [22]．

ここで一つの疑問が生ずる．関数 $g(y,z)$ がどのような形であれば，これから計算される拡張事後密度 $f_{x^n}(\theta^k)$ が漸近正規性を満たすか，という問題である．次節では，漸近正規性が成り立つために関数 $g(y,z)$ が満たすべき条件を示す．

3.2 主 定 理

まず，次の情報行列 $I^*(\theta^k)$ と $J^*(\theta^k)$ を定義しておく．

$$I^*(\theta^k)=E^*\left[\frac{\partial^2 L(d_{\theta^k}:X)}{\partial\theta^k(\partial\theta^k)^T}\right] \tag{29}$$

$$J^*(\theta^k)=E^*\left[\frac{\partial L(d_{\theta^k}:X)}{\partial\theta^k}\frac{\partial L(d_{\theta^k}:X)}{(\partial\theta^k)^T}\right] \tag{30}$$

$I^*(\theta^k)$ と $J^*(\theta^k)$ は常に存在するとは限らないが，本論文では $\forall\theta^k\in\Theta^k$ に対しこれらが存在するものとする．

損失最小推定量 $\hat{\theta}^k$ と真の分布に対する最適パラメータ θ^{k*} を

$$\hat{\theta}^k=\arg\min_{\theta^k}\frac{1}{n}\sum_{t=1}^{n}L(d_{\theta^k}:x_t) \tag{31}$$

$$\theta^{k*}=\arg\min_{\theta^k}E^*\left[L(d_{\theta^k}:X)\right] \tag{32}$$

とする．$\forall\delta>0$ に対し，球 $B_\delta(\theta^{k*})$ を $B_\delta(\theta^{k*})=\left\{\theta^k\in\Theta^k\middle|\|\theta^k-\theta^{k*}\|<\delta\right\}$ と定義する．同様に，$B_\delta(\tilde{\theta}^k)=\left\{\theta^k\in\Theta^k\middle|\|\theta^k-\tilde{\theta}^k\|<\delta\right\}$ としよう．

次に我々の解析に必要となる条件を示す．

[条件 1] (1) Θ^k はコンパクト集合である．

(2) $\exists c_1<\inf_{\theta^k\in\Theta^k}\pi(\theta^k)$，かつ $\sup_{\theta^k\in\Theta^k}\pi(\theta^k)<\exists c_2$ である．ただし，c_1 と c_2 はある正定数である．更に，$\pi(\theta^k)$ は Θ^k 上で連続 2 回微分可能である．

(3) θ^{k*} は Θ^k の内部で一意に定まる．

(4) 非負関数 $g(y,z)$ は，$y,z\in[0,\infty)$ である y に関して単調減少，z に関して単調増加関数である．また $g(y,z)$ は連続 2 回微分可能である．

(5) $I^*(\theta^k)$ は Θ^k 内の任意の θ^k に関して連続微分可能，かつ $I^*(\theta^{k*})$ は正定値である．

(6) $\forall n\in\{1,2,\cdots\}$ に対し $\frac{1}{n}\sum_{t=1}^{n}L(d_{\theta^k}:x_t)$ は，ほとんど確実に Θ^k 上で連続である．

(7) $E^*\left[L(d_{\theta^k}:X)\right]$ は Θ^k 上で任意の θ^k に関して連続 2 回微分可能である．$\forall n\in\{1,2,\cdots\}$ に対し $\frac{1}{n}\sum_{t=1}^{n}L(d_{\theta^k}:x_t)$ は，ほとんど確実に Θ^k 上で連続 2 回微分可能である．

(8) Θ^k 上で一様に

$$\frac{1}{n}\sum_{t=1}^{n}L(d_{\theta^k}:x_t)\to E^*\left[L(d_{\theta^k}:X)\right],\quad a.s.$$

$$\frac{1}{n}\frac{\partial}{\partial\theta^k}\sum_{t=1}^{n}L(d_{\theta^k}:x_t)\to\frac{\partial}{\partial\theta^k}E^*\left[L(d_{\theta^k}:X)\right],\quad a.s.$$

$$\frac{1}{n}\frac{\partial^2}{\partial\theta^k(\partial\theta^k)^T}\sum_{t=1}^{n}L(d_{\theta^k}:x_t)\to I^*(\theta^k),\quad a.s.$$

(9) (中心極限定理) $\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{\partial L(d_{\theta^{k*}}:X_t)}{\partial\theta^k}$ は正規分布 $N\left(\mathbf{0},J^*(\theta^{k*})\right)$ に法則収束する． □

この条件は，先の論文[9]の仮定と同一のものであり，パラメータ空間の境界などの特異点を除けば，広く成り立つ現実的な仮定といえる．同様の条件は[12], p.238で扱われている．この条件を満足する例については[9], [12]を参照．

もし条件(c.1')，(c.2')，(c.3')が成り立てば，拡張事後密度(21)は漸近正規性を満たす．以下では，条件1のもとで，条件(c.1')，(c.2')，(c.3')が成り立つための関数 $g(y,z)$ の条件を議論しよう．

[定理 1][注5] 条件1(1) ～ (8)を仮定する．

もし $y \to \infty$ のとき $\log g(y,z) \neq O\left(\log \frac{1}{y}\right)$ であれば，

$$f_{x^n}(\tilde{\theta}^k)\left(\det \Sigma_{x^n}\right)^{1/2} \to (2\pi)^{-k/2}, \quad a.s. \tag{33}$$

$$E^*\left[f_{x^n}(\tilde{\theta}^k)\left(\det \Sigma_{x^n}\right)^{1/2}\right] \to (2\pi)^{-k/2} \tag{34}$$

が成り立つ．漸近式(34)は条件1(8)の概収束を $p^*(\cdot)$ に対する平均収束に置き換えても成り立つ．

もし，条件1(8)が確率収束で置き換えられたとすると，

$$f_{x^n}(\tilde{\theta}^k)\left(\det \Sigma_{x^n}\right)^{1/2} \to (2\pi)^{-k/2}, \quad in\ prob. \tag{35}$$

となる．

一方，もし $y \to \infty$ のとき $\log g(y,z) = O\left(\log \frac{1}{y}\right)$ であれば，$f_{x^n}(\tilde{\theta}^k)$ の漸近正規性は成り立たない．

(証明) 付録参照． □

次に一様収束を議論するために，$\tilde{\theta}^k \in \overline{\Theta}^k$，かつ

$$\frac{1}{n}\sum_{t=1}^{n} L(d_{\tilde{\theta}^k} : x_t) < C \tag{36}$$

を満たす $x^n \in \mathcal{X}^n$ の部分集合 $\overline{\mathcal{X}}_C^n \subset \mathcal{X}^n$ を定義しよう．ここで $\overline{\Theta}^k$ は Θ^k の内部に含まれるある部分集合である．条件1(8)の代わりに，$x^n \in \overline{\mathcal{X}}_C^n$ となる x^n を考える．

更に，条件1(6)，(7)を次のように置き換える．

[条件 2] (6') $\forall x^n \in \overline{\mathcal{X}}_C^n$ と $\forall n \in \{1,2,\cdots\}$ に対し，$\frac{1}{n}\sum_{t=1}^{n} L(d_{\theta^k} : x_t)$ は Θ^k に関して連続である．

(7') $\forall x^n \in \overline{\mathcal{X}}_C^n$ と $\forall n \in \{1,2,\cdots\}$ に対し，$\frac{1}{n}\sum_{t=1}^{n} L(d_{\theta^k} : x_t)$ は Θ^k 上で2回連続微分可能である． □

[定理 2] 条件1(1)，(2)，(4)と条件2を仮定する．

もし $y \to \infty$ のとき $\log g(y,z) \neq O\left(\log \frac{1}{y}\right)$ であれば，

$$f_{x^n}(\tilde{\theta}^k)\left(\det \Sigma_{x^n}\right)^{1/2} \to (2\pi)^{-k/2} \tag{37}$$

が $x^n \in \overline{\mathcal{X}}_C^n$ に対して一様に成り立つ． □

(概証明) 条件2(5')と $x^n \in \overline{\mathcal{X}}_C^n$ より，$\frac{1}{n}\sum_{t=1}^{n} L''(d_{\tilde{\theta}^k} : x_t)$ と $\frac{1}{n}\sum_{t=1}^{n} L''(d_{\tilde{\theta}^k} : x_t)$ は有界である．したがって，定理1の証明と同様の展開により，定理が得られる[13]． □

$\log g(y,z) \neq O\left(\log \frac{1}{y}\right)$ という条件は，$g(y,z)$ が y に関して多項式オーダの減少関数では漸近正規性は成り立たず，指数オーダ以上になると成り立つことを示している．この $g(y,z)$ の条件より，式(23)の拡張事後密度は漸近正規性を有さないことがわかる．同様に，式(24)も漸近正規性を満たさない．一方，式(25)は漸近正規性を満たす．しかし定理1より，$g(y,z)$ が y に関して指数オーダより早く減少するようなものであっても，漸近正規性を満たすこともわかる．

4. 拡張確率的コンプレキシティの評価

与えられた正定数 $\lambda > 0$ のもとで．拡張確率的コンプレキシティ(ESC)は次式で定義される[23], [24].

$$ESC(x^n) = -\frac{1}{\lambda}\log \int \exp\left(-\lambda \sum_{t=1}^{n} L(d_{\theta^k} : x_t)\right)\pi(\theta^k)d\theta^k \tag{38}$$

ESCはオンライン学習のthe aggregating algorithmに適用されるなど，学習問題の性能を保証するために有用であることが示されている[23]．その精密な漸近式を求めることは，ESCを用いた学習アルゴリズムの漸近性能に対する保証を与える．

一方，本論文で与えた拡張事後密度の漸近正規性を応用することにより，ESCの漸近式が自然に導かれる[9]．概収束と平均収束の意味でのESCの評価は[9]で行っているが，ここでは一様収束の結果も示し，前節の漸近正規性の成立条件と併せて考察する．これにより，ESCの漸近式において漸近正規性が本質的であるだけでなく，ESCが漸近正規性という観点からも自然なものであることが再認識される．

まず，$h(x^n)$, $h(z^n|\theta^k)$, $\pi(\theta^k|x^n)$ を

$$h(x^n) = \int \exp\left(-\lambda \sum_{t=1}^{n} L(d_{\theta^k} : x_t)\right)\pi(\theta^k)d\theta^k \tag{39}$$

(注5)：本論文では拡張事後密度最大推定量 $\tilde{\theta}^k$ を用いた展開のみを示す．$\tilde{\theta}^k$ を $\hat{\theta}^k$ で置き換えても，[9]と同様の議論により全く同様の結果が成り立つことが容易にわかる．

$$h(x^n|\theta^k) = \exp\left(-\lambda \sum_{t=1}^{n} L(d_{\theta^k} : x_t)\right) \tag{40}$$

$$\pi(\theta^k|x^n) = \frac{\exp\left(-\lambda \sum_{t=1}^{n} L(d_{\theta^k} : x_t)\right)\pi(\theta^k)}{\int \exp\left(-\lambda \sum_{t=1}^{n} L(d_{\theta^k} : x_t)\right)\pi(\theta^k)d\theta^k} \tag{41}$$

と定義する．ただし，$\pi(\theta^k|x^0) = \pi(\theta^k)$, $L(d_{\theta^k} : x_0) = 0$ である．

このとき，次の補題が成り立つ．

[補題 3] [9] $h(x^n)$, $h(x^n|\theta^k)$, $\pi(\theta^k|x^n)$, に対し次式の関係が成り立つ．

$$h(x^n) = \frac{h(x^n|\theta^k)\pi(\theta^k)}{\pi(\theta^k|x^n)} \tag{42}$$

□

式 (42) はベイズ規則

$$p(x^n) = \frac{p(x^n|\theta^k)\pi(\theta^k)}{f_B(\theta^k|x^n)} \tag{43}$$

の一般化である．式 (42) より，

$$ESC(x^n) = -\frac{1}{\lambda}\log h(x^n|\theta^k)\pi(\theta^k) + \frac{1}{\lambda}\log \pi(\theta^k|x^n) \tag{44}$$

となるから，$\log \pi(\theta^k|x^n)$ の漸近式が与えられれば，$ESC(x^n)$ の漸近式が得られることがわかる．

$\pi(\theta^k|x^n)$ は Θ^k 上の密度関数であるが，一般にはベイズ規則で計算されるベイズ事後確率密度ではない．もし対数損失と $\lambda = 1$ を適用すると，$\pi(\theta^k|x^n)$ はベイズ事後確率密度となり，ESC は確率的コンプレキシティ (SC) [17]

$$SC(x^n) = -\log \int_{\theta^k} p_{\theta^k}(x^n)\pi(\theta^k)d\theta^k \tag{45}$$

となる．

定理 1 の漸近正規性より，次の定理が得られる [9].

[定理 3] [9] 条件 1(1)～(8) のもとで，

$$\frac{\pi(\tilde{\theta}^k|x^n)}{\sqrt{n}^k} \to \left(\frac{\lambda}{2\pi}\right)^{k/2}\sqrt{\det I^*(\tilde{\theta}^k)}, \quad a.s. \tag{46}$$

が成り立つ． □

また，$\tilde{\theta}^k$ は $\hat{\theta}^k$ で置き換えることもできる [9]．定理 3 は，

$$(\Sigma_{x^n})^{-1} = -\lambda \sum_{t=1}^{n} \frac{\partial L(d_{\theta^k} : x_t)}{\partial\theta^k(\partial\theta^k)^T} + \frac{\partial^2 \log \pi(\theta^k)}{\partial\theta^k(\partial\theta^k)^T} \tag{47}$$

より得られる

$$-\frac{1}{n}(\Sigma_{x^n})^{-1} \to \lambda I^*(\tilde{\theta^k}), \quad a.s. \tag{48}$$

という式から導かれる．

この定理と補題 3 より，直ちに次の定理が得られる．

[定理 4] [9] 条件 1(1)～(8) のもとで，$ESC(x^n)$ の漸近式は

$$ESC(x^n) = \sum_{t=1}^{n} L(d_{\hat{\theta}^k} : x_t) + \frac{k}{2\lambda}\log\frac{n\lambda}{2\pi} + \frac{1}{\lambda}\log\frac{\sqrt{\det I^*(\hat{\theta}^k)}}{\pi(\hat{\theta}^k)} + o(1), \quad a.s. \tag{49}$$

$$= \sum_{t=1}^{n} L(d_{\tilde{\theta}^k} : x_t) + \frac{k}{2\lambda}\log\frac{n\lambda}{2\pi} + \frac{1}{\lambda}\log\frac{\sqrt{\det I^*(\tilde{\theta}^k)}}{\pi(\tilde{\theta}^k)} + o(1), \quad a.s. \tag{50}$$

と与えられる． □

上の定理は真の分布に対する概収束の意味で導かれている．すなわち，$\hat{\theta}^k$ と $\tilde{\theta}^k$ は確率変数である．

[補題 4] [12], pp.240-241 条件 1のもとで，

$$E^*\left[\frac{\sum_{t=1}^{n} L(d_{\hat{\theta}^k} : X_t)}{\sum_{t=1}^{n} L(d_{\theta^{k*}} : X_t)}\right] \to \frac{TrJ^*(\theta^{k*})\{I^*(\theta^{k*})\}^{-1}}{2} \tag{51}$$

が成り立つ．ただし，Tr は行列のトレースを表す．

□

補題 4 より，$ESC(x^n)$ の期待値 $E^*[ESC(X^n)]$ の漸近評価が可能となる．

[定理 5] [9] 条件 1 のもとで，$E^*[ESC(X^n)]$ は

$$E^*[ESC(X^n)] = E^*[L(d_{\theta^{k*}} : X_t)] + \frac{k}{2\lambda}\log\frac{n\lambda}{2\pi}$$

$$-\frac{TrJ^*(\theta^{k*})\{I^*(\theta^{k*})\}^{-1}}{2\lambda}$$
$$+\frac{1}{\lambda}\log\frac{\sqrt{\det I^*(\theta^{k*})}}{\pi(\theta^{k*})}+o(1) \quad (52)$$

を満たす. □

式 (49) と式 (52) より，SC の漸近式が

$$SC(x^n)=-\log p_{\hat{\theta}^k}(x^n)+\frac{k}{2}\log\frac{n}{2\pi}$$
$$+\log\frac{\sqrt{\det I^*(\hat{\theta}^k)}}{\pi(\hat{\theta}^k)}+o(1), \quad a.s. \quad (53)$$

$$E^*\left[SC(X^n)\right]$$
$$=-E^*\left[\log p_{\theta^{k*}}(X^n)\right]+\frac{k}{2}\log\frac{n}{2\pi}$$
$$-\frac{TrJ^*(\theta^{k*})\{I^*(\theta^{k*})\}^{-1}}{2}+\log\frac{\sqrt{\det I^*(\theta^{k*})}}{\pi(\theta^{k*})}$$
$$+o(1) \quad (54)$$

と与えられることもわかる.

以上の結果は，概収束と平均収束の意味での ESC の漸近式である．これに対し Yamanishi は平均的意味とともに，データ系列に対して一様に成り立つ ESC の漸近的上界を導いている [23] が，一様に成り立つ上界を考える代償として，厳しい漸近式は得られていない．平均収束の意味での ESC の漸近式 (52) は Yamanishi の上界がきついものであることを示しており，概収束の意味での ESC の漸近式 (49) は同様の収束がほとんど確実に得られる系列に対して成り立つことを述べている．

一方，データ系列に対して一様に漸近正規性が成り立つかどうかを考えると，一般に答えは否であり，データ系列の部分集合を考えなければならない．これに対する結果は以下のようになる．

[定理 6] 条件 1(1), (2), (4) と条件 2 のもとで,

$$ESC(x^n)=\sum_{t=1}^{n}L(d_{\tilde{\theta}^k}:x_t)+\frac{k}{2\lambda}\log\frac{n\lambda}{2\pi}$$
$$+\frac{1}{\lambda}\log\frac{\sqrt{\det\hat{I}(\tilde{\theta}^k)}}{\pi(\hat{\theta}^k)}+o(1), \quad (55)$$

が $x^n\in\overline{\mathcal{X}}_C^n$ に対して一様に成り立つ．ここで $\hat{I}(\theta^k)$ は経験情報行列であり，

$$\hat{I}(\theta^k)=\frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2 L(d_{\theta^k}:x_t)}{\partial\theta^k(\partial\theta^k)^T} \quad (56)$$

で与えられる. □

以上の結果は，ESC の漸近式の定数オーダの項が本質的に漸近正規性から導かれることを示している．また，定理 1 から，式 (41) の拡張事後密度は漸近正規性を満たすための条件の境界に位置するものであることがわかる．条件 1 のもとで，損失最小推定量を基準化した統計量は漸近的に正規分布 $N\left(\mathbf{0},\{I^*(\theta^k)\}^{-1}J^*(\theta^k)\{I^*(\theta^k)\}^{-1}\right)$ に法則収束することが知られているが [12]，事後密度の共分散がこの $\{I^*(\theta^k)\}^{-1}J^*(\theta^k)\{I^*(\theta^k)\}^{-1}$ に比べて過度に小さいのは意味がない．なぜなら，共分散 $\{I^*(\theta^k)\}^{-1}J^*(\theta^k)\{I^*(\theta^k)\}^{-1}$ 以上にパラメータを精度良く特定するための情報は，確率的に生起するデータ系列には含まれていないからである．逆に関数 $g(y,z)$ を調整して事後密度の共分散を大きくしようとすると，いずれ漸近正規性が成り立たなくなる．それ以上共分散のオーダを大きくすると漸近正規性を満たさなくなるような境界に位置する場合が ESC で定義されている拡張事後密度であり，拡張事後密度の共分散と損失最小推定量の共分散が同じオーダになるようなケースになっている．その意味からも ESC は自然なものであることが再認識できる.

5. むすび

本論文では，損失関数を考慮した拡張事後密度による推定を考え，その漸近正規性の成り立つ条件を示した．更に，この結果を用いて確率的コンプレキシティ，拡張確率的コンプレキシティの漸近式を示した.

漸近正規性は，総計的推測において本質的な性質である．しかし，拡張確率的コンプレキシティの評価においては，個々のデータ系列 $\mathcal{X}^n$ に対して一様に成り立つ漸近式が重要視されている．漸近正規性から考察した場合，パラメータの境界領域などの経験情報行列 $\hat{I}(\tilde{\theta}^k)$ が発散する点において，一様性が保証できない．このようなケースの考察のためには，異なる視点が解析が必要である [20].

文　献

[1] J.M. Bernardo and A.F.M. Smith, Bayesian Theory, John Wiley & Sons, 1994.

[2] G.E.P. Box and G.C. Tiao, Bayesian Inference in Statistical Analysis, John Wiley & Sons, Inc., 1992.

[3] B.S. Clarke, "Asymptotic normality of the posterior in relative entropy," IEEE Trans. Inf. Theory, vol.45, no.1, pp.165–176, 1999.

[4] B.S. Clarke and A.R. Barron, "Information—Theoretic asymptotics of Bayes methods," IEEE

Trans. Inf. Theory, vol.36, no.3, pp.453–471, 1990.
[5] C.-F. Chen, "On asymptotic normality of limiting density function with Bayesian implications," J. R. Statist. Soc. B, vol.47, no.3, pp.540–546, 1988.
[6] W. Feller, An Introduction to Probability and Its Applications, vol.1 and 2, John Wiley & Sons, New York, 1957.
[7] T.S. Ferguson, Mathematical Statistics—A Decision Theoretic Approach, New York and London, Academic, 1967.
[8] M. Gotoh, T. Matsushima, and S. Hirasawa, "A generalization of B.S. Clarke and A.R. Barron's asymptotics of Bayes codes for FSMX sources," IEICE Trans. Fundamentals, vol.E81-A, no.10, pp.2123–2132, 1998.
[9] M. Gotoh, T. Matsushima, and S. Hirasawa, "Almost sure and mean convergence of extended stochastic complexity," IEICE Trans. Fundamentals, vol.E82-A, no.10, pp.2129–2137, Oct. 1999.
[10] J.A. Hartigan, Bayes Theory, Springer-Verlag, 1983.
[11] C.C. Heyde and I.M. Johnstone, "On asymptotic posterior normality for stochastic process," J. R. Statist. Soc. B, vol.41, no.2, pp.184–189, 1979.
[12] H. Linhart and W. Zucchini, Model Selection, John Wiley & Sons, 1986.
[13] 丸山英昭，後藤正幸，平澤茂一，"事後確率密度の漸近正規性に関する一考察，" 信学技報，IT99-28, 1999.
[14] N. Murata, S. Yoshizawa, and S. Amari, "Network information criterion—Determining the number of hidden units for an artificial neural network model," IEEE Trans. Neural Networks, vol.5, no.6, pp.865–872, 1994.
[15] D.S. Poskitt, "Precision, complexity and Bayesian model determination," J. R. Statist. Soc. B, vol.49, no.2, pp.199–208, 1987.
[16] J. Rissanen, "Universal conding, information, prediction, and estimation," IEEE Trans. Inf. Theory, vol.IT-30, no.4, pp.629–636, 1984.
[17] J. Rissanen, "Stochastic complexity," J. R. Statist. Soc. B, vol.49, pp.223–265, 1987.
[18] J. Rissanen, "Fisher information and stochastic complexity," IEEE Trans. Inf. Theory, vol.42, no.1, pp.40–47, 1996.
[19] 竹内　啓，"情報量基準の分布とモデルの適切さの規準，" 数理科学，no.153, pp.12–18, 1976.
[20] 竹内純一，"確率的コンプレキシティと Jeffreys 混合予測戦略，" 1998 年情報論的学習理論ワークショップ予稿集，pp.9–16, 1998.
[21] A.M. Walker, "On the asymptotic behaviour of posterior distributions," J. R. Statist. Soc. B, vol.31, pp.80–88, 1969.
[22] 渡辺澄夫，"ベイズ法による階層型統計モデルの推定誤差について，" 信学論（A），vol.J81-A, no.10, pp.1442–1452, Oct. 1998.
[23] K. Yamanishi, "A decision-theoretic extension of stochastic complexity and its applications to learning," IEEE Trans. Inf. Theory, vol.44, no.4, pp.1424–1439, 1998.
[24] 山西健司，"拡張型確率的コンプレキシティと学習理論，" 1998 年情報論的学習理論ワークショップ予稿集，pp.33–40, 1998.

付　　録

定理 1 の証明

紙面の都合上，概収束に関する証明のみを与える．他は収束の強さの違いであるが，ほぼ同様に証明できる．

漸近正規性を示すためには，補題 2 の条件 (c.1')～(c.3') が成り立つことを示せばよい．

簡単のため，

$$g'(y,z) = \frac{\partial g(y,z)}{\partial y} \tag{A·1}$$

$$g''(y,z) = \frac{\partial^2 g(y,z)}{(\partial y)^2} \tag{A·2}$$

$$g = g\left(\sum_{t=1}^{n} L(d_{\theta^k} : x_t), \pi(\theta^k)\right) \tag{A·3}$$

$$g' = g'\left(\sum_{t=1}^{n} L(d_{\theta^k} : x_t), \pi(\theta^k)\right) \tag{A·4}$$

$$g'' = g''\left(\sum_{t=1}^{n} L(d_{\theta^k} : x_t), \pi(\theta^k)\right) \tag{A·5}$$

などと記述する．条件 1(4), (7) から

$$\frac{\partial}{\partial \theta^k} K_{x^n}(\theta^k) = \frac{\partial \log g}{\partial \theta^k} = \frac{1}{g}\frac{\partial g}{\partial \theta^k} \tag{A·6}$$

$$\frac{\partial^2 K_{x^n}(\theta^k)}{\partial \theta^k (\partial \theta^k)^T} = \frac{1}{g}\frac{\partial^2 g}{\partial \theta^k (\partial \theta^k)^T} - \frac{1}{g^2}\frac{\partial g}{\partial \theta^k}\frac{\partial g}{(\partial \theta^k)^T} \tag{A·7}$$

が存在する．ここで，

$$\frac{\partial g}{\partial \theta^k} = g' \sum_{t=1}^{n} \frac{\partial L(d_{\theta^k} : x_t)}{\partial \theta^k} \tag{A·8}$$

$$\begin{aligned}\frac{\partial^2 g}{\partial \theta^k (\partial \theta^k)^T} &= g'' \sum_{t=1}^{n} \frac{\partial L(d_{\theta^k} : x_t)}{\partial \theta^k} \sum_{t=1}^{n} \frac{\partial L(d_{\theta^k} : x_t)}{(\partial \theta^k)^T} \\ &\quad + g' \sum_{t=1}^{n} \frac{\partial^2 L(d_{\theta^k} : x_t)}{\partial \theta^k (\partial \theta^k)^T}\end{aligned} \tag{A·9}$$

であるから，

$$\frac{\partial^2 K_{x^n}(\theta^k)}{\partial\theta^k(\partial\theta^k)^T} = \frac{g'}{g}\sum_{t=1}^{n}\frac{\partial^2 L(d_{\theta^k}:x_t)}{\partial\theta^k(\partial\theta^k)^T} + \frac{g''g-(g')^2}{g^2}\sum_{t=1}^{n}\frac{\partial L(d_{\theta^k}:x_t)}{\partial\theta^k}\sum_{t=1}^{n}\frac{\partial L(d_{\theta^k}:x_t)}{(\partial\theta^k)^T} \tag{A·10}$$

が成り立つ．

一方，条件 (c.1') は

$$\det\left\{\frac{\partial^2 K_{x^n}(\theta^k)}{\partial\theta^k(\partial\theta^k)^T}\right\}\to\infty, \quad a.s. \tag{A·11}$$

を意味する．そこで式 (A·10) から式 (A·11) の成り立つための条件を考える．

まず，条件 1(8) より，

$$\sum_{t=1}^{n}\frac{\partial L(d_{\theta^k}:x_t)}{\partial\theta^k}\sum_{t=1}^{n}\frac{\partial L(d_{\theta^k}:x_t)}{\partial\theta^k} = O(n^2), \quad a.s. \tag{A·12}$$

$$\sum_{t=1}^{n}\frac{\partial^2 L(d_{\theta^k}:x_t)}{\partial\theta^k(\partial\theta^k)^T} = O(n), \quad a.s. \tag{A·13}$$

である．また，g, g', g'' の中身である $\sum_{t=1}^{n} L(d_{\theta^k}:x_t)$ を考えても同様の議論から

$$\sum_{t=1}^{n} L(d_{\theta^k}:x_t) = O(n), \quad a.s. \tag{A·14}$$

が成り立つ．式 (A·10) より，式 (A·11) の成り立つための条件は，

$$\frac{g'}{g}\sum_{t=1}^{n}\frac{\partial^2 L(d_{\theta^k}:x_t)}{\partial\theta^k(\partial\theta^k)^T}\to\infty, \quad a.s.$$

または

$$\frac{g''g-(g')^2}{g^2}\sum_{t=1}^{n}\frac{\partial L(d_{\theta^k}:x_t)}{\partial\theta^k}\sum_{t=1}^{n}\frac{\partial L(d_{\theta^k}:x_t)}{(\partial\theta^k)^T}\to\infty, \quad a.s.$$

が成り立たなければならない．したがって，式 (A·12)〜(A·14) と，$g(y,z)$ が y に関して単調減少であることから，(c.1') が成り立つためには，$y\to\infty$ のとき

$$\left|\frac{g''(y,z)}{g(y,z)}y\right| = \left|\frac{y}{g(y,z)}\frac{\partial^2 g(y,z)}{(\partial y)^2}\right|\to\infty \tag{A·15}$$

$$\left|\frac{g'(y,z)}{g(y,z)}y\right| = \left|\frac{y}{g(y,z)}\frac{\partial g(y,z)}{\partial y}\right|\to\infty \tag{A·16}$$

であることが必要十分である．したがって，$y\to\infty$ のとき $\log g(y,z)\neq O\left(\log\frac{1}{y}\right)$ でなければならない．逆にもし $\log g(y,z)=O\left(\log\frac{1}{y}\right)$ であれば，(c.1') は成り立たないので漸近正規性も成り立たないことがわかる．この条件は，$g(y,z)$ が y に関して多項式オーダの減少関数では (c.1') は成り立たず，指数オーダ以上になると (c.1') が成り立つこと，すなわち任意の正定数 $\gamma>0$ に対して $g(y,z)\neq O\left(n^{-\gamma}\right)$ のとき (c.1') が成り立つことを示している．一方，このとき条件 1(5),(8) より，(c.2') も成り立つ．

ここで，条件 (c.1'),(c.2') のもとでは，

$$\lim_{n\to\infty} f_{x^n}(\tilde{\theta}^k)(\det\Sigma_{x^n})^{1/2}\leqq(2\pi)^{-k/2}, \quad a.s. \tag{A·17}$$

が成り立ち，更に (c.3') も同時に成り立てば，

$$\lim_{n\to\infty} f_{x^n}(\tilde{\theta}^k)(\det\Sigma_{x^n})^{1/2}\to(2\pi)^{-k/2}, \quad a.s. \tag{A·18}$$

が成り立つ [1].

よって，最後に $\log g(y,z)\neq O\left(\log\frac{1}{y}\right)$, $(y\to\infty)$ と式 (A·17) から，(c.3') が成り立つことを示そう．そのために

$$\begin{aligned} &K_{x^n}(\theta^k)-K_{x^n}(\theta^{k*})\\ &\quad=\log\frac{f_{x^n}(\theta^k)}{f_{x^n}(\theta^{k*})}\\ &\quad=\log\frac{g\left(\sum_{t=1}^{n}L(d_{\theta^k}:x_t),\pi(\theta^k)\right)}{g\left(\sum_{t=1}^{n}L(d_{\theta^{k*}}:x_t),\pi(\theta^{k*})\right)} \end{aligned} \tag{A·19}$$

を評価したい．

$$\begin{aligned} &\frac{1}{\beta_{\theta^k}(x^n)}\log\frac{g\left(\sum_{t=1}^{n}L(d_{\theta^k}:x_t),\pi(\theta^k)\right)}{g\left(\sum_{t=1}^{n}L(d_{\theta^{k*}}:x_t),\pi(\theta^{k*})\right)}\\ &\quad=\log\frac{g\left(\frac{1}{n}\sum_{t=1}^{n}L(d_{\theta^k}:x_t),\pi(\theta^k)\right)}{g\left(\frac{1}{n}\sum_{t=1}^{n}L(d_{\theta^{k*}}:x_t),\pi(\theta^{k*})\right)}, \quad a.s. \end{aligned} \tag{A·20}$$

となるような x^n の関数 $\beta_{\theta^k}(x^n)$ を考える．条件 1(4), (8) から，$\forall\theta^k\in\Theta^k$ に対し一様に，

$$\begin{aligned} &g\left(\frac{1}{n}\sum_{t=1}^{n}L(d_{\theta^k}:x_t),\pi(\theta^k)\right)\\ &\quad\to g\left(E^*[L(d_{\theta^k}:X)],\pi(\theta^k)\right), \quad a.s.\\ &\quad=O(1), \quad a.s. \end{aligned} \tag{A·21}$$

$$\sum_{t=1}^{n} L(d_{\theta^k} : x_t) = O(n), \quad a.s. \tag{A·22}$$

が成り立つ．よって，$\beta_{\theta^k}(x^n)$ は漸近的にはほとんど確実に x^n に依存せず，n だけに依存する関数に置き換えることができる．したがって，$n \to \infty$ のとき

$$\frac{1}{\beta_{\theta^k}(n)} \log \frac{g\left(\sum_{t=1}^{n} L(d_{\theta^k} : x_t), \pi(\theta^k)\right)}{g\left(\sum_{t=1}^{n} L(d_{\theta^{k*}} : x_t), \pi(\theta^{k*})\right)} = \log \frac{g\left(\frac{1}{n}\sum_{t=1}^{n} L(d_{\theta^k} : x_t), \pi(\theta^k)\right)}{g\left(\frac{1}{n}\sum_{t=1}^{n} L(d_{\theta^{k*}} : x_t), \pi(\theta^{k*})\right)}, \quad a.s. \tag{A·23}$$

となるような $\beta_{\theta^k}(n)$ が存在する．一方，

$$\frac{1}{n}\left\{\sum_{t=1}^{n} L(d_{\theta^k} : x_t) - \sum_{t=1}^{n} L(d_{\theta^{k*}} : x_t)\right\} \to E^*[L(d_{\theta^k} : X)] - E^*[L(d_{\theta^{k*}} : X)], \quad a.s. \tag{A·24}$$

であり，$\forall\delta > 0$, $\forall\theta^k \notin B_\delta(\theta^{k*})$ に対し $E^*[L(d_{\theta^k} : X)] - E^*[L(d_{\theta^{k*}} : X)] > C$ となる正定数 $C > 0$ が存在するから，

$$\sum_{t=1}^{n} L(d_{\theta^k} : x_t) - \sum_{t=1}^{n} L(d_{\theta^{k*}} : x_t) = O(n), \quad a.s. \to \infty, \quad a.s. \tag{A·25}$$

が成り立つ．したがって，式 (A·22)，(A·25) と $\log g(y,z) \neq O\left(\log \frac{1}{y}\right)$ $(y \to \infty)$ であることから，$\forall\delta > 0$, $\forall\theta^k \notin B_\delta(\theta^{k*})$ に対して

$$\log \frac{g\left(\sum_{t=1}^{n} L(d_{\theta^k} : x_t), \pi(\theta^k)\right)}{g\left(\sum_{t=1}^{n} L(d_{\theta^{k*}} : x_t), \pi(\theta^{k*})\right)} \neq O\left(\log \frac{1}{n}\right)$$

$$\log \frac{g\left(\sum_{t=1}^{n} L(d_{\theta^k} : x_t), \pi(\theta^k)\right)}{g\left(\sum_{t=1}^{n} L(d_{\theta^{k*}} : x_t), \pi(\theta^{k*})\right)} \to -\infty, \quad a.s. \tag{A·26}$$

となる(注6)．一方，式 (A·23) の右辺は，条件 1(8) より，θ^k に関して一様に

$$\log \frac{g\left(\frac{1}{n}\sum_{t=1}^{n} L(d_{\theta^k} : x_t), \pi(\theta^k)\right)}{g\left(\frac{1}{n}\sum_{t=1}^{n} L(d_{\theta^{k*}} : x_t), \pi(\theta^{k*})\right)} \to \log \frac{g\left(E^*[L(d_{\theta^k} : X)], \pi(\theta^k)\right)}{g\left(E^*[L(d_{\theta^{k*}} : X)], \pi(\theta^{k*})\right)} < 0, \quad a.s.$$

と収束するから，$\forall\theta^k \notin B_\delta(\theta^{k*})$ に対して $\beta_{\theta^k}(n) \neq O(\log n)$, $\beta_{\theta^k}(n) \to \infty$ が成り立つ．これは，任意の $\gamma > 0$ に対して $\exp\{\beta_{\theta^k}(n)\} \neq O(n^\gamma)$ であることを意味する．同時に，$\forall\delta > 0$ に対して正定数 $\exists C_\delta > 0$ が存在し，$n \to \infty$ のとき $\forall\theta^k \notin B_\delta(\theta^{k*})$ に対して一様に

$$\frac{1}{\beta_{\theta^k}(n)} \log \frac{f_{x^n}(\theta^k)}{f_{x^n}(\theta^{k*})} < -C_\delta, \quad a.s. \tag{A·27}$$

が成り立つこともわかる．

$\beta(n)$ を $\beta(n) = \inf_{\theta^k \notin B_\delta(\theta^{k*})} \beta_{\theta^k}(n)$ のように定めると，$\forall\delta > 0$ に対し，$n \to \infty$ のとき，$\forall\theta^k \notin B_\delta(\theta^{k*})$ に対して一様に

$$\frac{f_{x^n}(\theta^k)}{f_{x^n}(\theta^{k*})} < \exp\{-\beta(n)C_\delta\}, \quad a.s. \tag{A·28}$$

が成り立つ．

一方，式 (A·17) から，$n \to \infty$ のとき，

$$f_{x^n}(\theta^{k*}) \leqq (2\pi)^{-k/2}\left(\det K''_{x^n}(\theta^k)\right)^{1/2}, \quad a.s. \tag{A·29}$$

が成り立っている．条件 1(8) と式 (A·10) より，

$$\det K''_{x^n}(\theta^k) = O\left(\max\left\{\left(\frac{ng'}{g}\right)^k, \left(\frac{ng''}{g}\right)^k\right\}\right) \quad a.s. \tag{A·30}$$

であるので，不等式 (A·28) から，$\forall\delta > 0$ に対し，$\theta^k \notin B_\delta(\theta^{k*})$ に対して一様に

$$f_{x^n}(\theta^k) < f_{x^n}(\theta^{k*})\exp\{-\beta(n)C_{\delta_\epsilon}\} = O\left(\frac{\max\left\{\left(\frac{ng'}{g}\right)^k, \left(\frac{ng''}{g}\right)^k\right\}}{\exp\{\beta(n)C_\delta\}}\right) \to 0, \quad a.s. \tag{A·31}$$

が成り立つ．最後の収束は，$\forall\gamma > 0$ に対し $\exp\{\beta_{\theta^k}(n)\} = \Omega(n^\gamma)$ である(注7)ことと，式 (22) の $\frac{g'}{g} = O(1)$ から明らかである．

条件 1(1) より，Θ^k はコンパクトであるから，$\forall\delta > 0$ に対し

(注6)：n の関数 $y_1(n)$, $y_2(n)$ が，$y_1(n) = O(n)$, $y_2(n) = O(n)$, $y_1(n) - y_2(n) = O(n)$ をみたしているとき，$n \to \infty$ において $\frac{\log g(n,z)}{\log \frac{1}{n}} \to \infty$ であれば，$\frac{\log g(y_1(n),z) - \log g(y_2(n),z)}{\log \frac{1}{n}} \to \infty$ が成り立つため．

(注7)：$\exp\{\beta_{\theta^k}(n)\} \neq O(n^\gamma)$ より明らかに $\exp\{\beta_{\theta^k}(n)\} \neq o(n^\gamma)$ であるから，$\exp\{\beta_{\theta^k}(n)\} = \Omega\left(n^\gamma\right)$ である．

$$\int_{\theta^k \notin B_\delta(\theta^{k*})} f_{x^n}(\theta^k)d\theta^k \to 0, \quad a.s. \qquad \text{(A·32)}$$

となり，これは

$$\int_{\theta^k \in B_\delta(\theta^{k*})} f_{x^n}(\theta^k)d\theta^k \to 1, \quad a.s. \qquad \text{(A·33)}$$

であることを意味する．

$\tilde{\theta}^k \to \theta^{k*}$ $a.s.$ であることから，$0 < \forall\delta' < \forall\delta''$ であれば，$n \to \infty$ のとき，$B_{\delta'}(\theta^{k*}) \subset B_{\delta''}(\hat{\theta}^k)$, $a.s.$ したがって (c.3') が成り立ち，式 (33) が証明された．

式 (34) は，式 (33) と有界収束定理から直ちに導かれる． □

（平成 11 年 10 月 22 日受付，12 年 1 月 11 日再受付）

後藤　正幸　（正員）

平 4 武蔵工大・工・経営卒．平 6 同大大学院修士課程了．平 6 早大・理工学研究科博士後期課程入学．平 8〜11 同大理工学部・経営システム工学科助手．平 11 同大メディアネットワークセンター等・非常勤講師．現在，東大大学院・工学系研究科・環境海洋工学専攻・助手．情報源符号化，統計的学習理論，統計的モデル選択，ベイズ統計応用などの研究に従事．平 12 より，ビジネスモデル，コストモデル等の研究にも着手．工博．IEEE, 情報理論とその応用学会，人工知能学会，日本経営工学会各会員．

松嶋　敏泰　（正員）

昭 53 早大・理工・工業経営卒．昭 55 同大大学院修士課程了．同年，日本電気（株）入社．昭 61 早大・理工学研究科博士後期課程入学．平 1 横浜商科大学講師．平 3 同大助教授．平 4 早大・理工学部・工業経営学科（現在経営システム工学科）助教授，現在教授．知識情報処理及び情報理論とその応用に関する研究に従事．工博．IEEE, 情報理論とその応用学会，人工知能学会，情報処理学会，OR 学会，日本経営工学会等各会員．

平澤　茂一　（正員）

昭 36 早大・理工・数学卒．昭 38 同電気通信卒．同年三菱電機（株）入社．昭 56 早大・理工・工業経営学科（現在経営システム工学科）教授，現在に至る．情報理論とその応用，データ伝送方式，並びに計算機応用システムの開発などの研究に従事．工博．昭 54UCLA 計算機科学科客員研究員．昭 60 ハンガリー科学アカデミー，昭 61 伊トリエステ大学客員研究員．平 5 電子情報通信学会 小林記念特別賞，業績賞受賞．平 8 情報理論とその応用学会会長．IEEE(Fellow)，情報理論とその応用学会，人工知能学会，情報処理学会，OR 学会，日本経営工学会等各会員．

論　文

ウェーブレットパケット基底を用いた信号推定におけるベイズ決定理論の適用に関する一考察

北原　正樹[†a)]　　野村　亮[††]　　松嶋　敏泰[††]

A Note on Signal Estimation by Wavelet Packets and Bayes Decision Theory

Masaki KITAHARA[†a)], Ryo NOMURA[††], and Toshiyasu MATSUSHIMA[††]

あらまし　雑音が混入した観測信号から未知の信号を推定する問題において，正規直交基底を用いた推定法が従来から研究されている．本研究では，ウェーブレットパケット基底の集合を用いて推定を行う場合の問題設定を扱う．従来においては，この問題設定においてベイズ的手法の適用が試みられているものの，未知信号と推定信号の間の2乗誤差を損失関数とした場合のベイズ最適な推定に関しては考察されていなかった．本研究では2乗誤差損失のもとでのベイズ最適な推定量，推定信号を効率的に計算するアルゴリズムを提案し，推定量の性能と性質に関して考察する．

キーワード　信号推定，ウェーブレットパケット基底，ベイズ決定理論，木構造モデル

1. まえがき

本研究では雑音が混入した観測信号から未知の信号を推定する問題を扱う．従来研究においては，観測信号に仮定されるモデルなど，問題の性質に応じた種々の推定法が研究されている [1]．そのような中で，直交ウェーブレット基底等の正規直交基底(注1)による未知信号の表現を推定に利用する研究が近年盛んに研究されている [3], [4], [6]〜[8]．

これらの研究の多くでは，観測信号は次式で与えられると仮定される．

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e} \tag{1}$$

$\boldsymbol{x} = (x(0), x(1), \cdots, x(N-1))^T$，$\boldsymbol{x} \in \boldsymbol{R}^N$は未知の離散時間信号を表し，$\boldsymbol{e} = (e(0), e(1), \cdots, e(N-1))^T$の要素はiidで$N(0, \sigma^2)$に従う白色雑音であり，$\sigma^2$は既知であるとする．この問題では，基底と未知信号の変換係数が決定することによって未知信号が一意に定まる．すなわち，この推定問題は基底と未知信号の変換係数を未知パラメータとした推定問題となる．

なお，従来において扱われていた問題設定は，大きく分けて以下に述べる二つ（問題設定A，B）に分類できる．問題設定Aにおいては，未知パラメータをなす基底があらかじめ一つに定められている [6], [7]．この場合，変換係数を未知としたもとでの推定問題になる．問題設定Bにおいては，基底は一つに定められていないものの，基底の有限集合が定められている [3], [4], [8]．この問題設定においては，基底（基底の集合に含まれるいずれかの基底）と変換係数を未知としたもとでの推定問題になる．本研究では後者の場合を扱い，未知信号と推定信号の間の2乗誤差をできるだけ小さくすることを目指した推定を考える．ただし，基底の集合はウェーブレットパケット基底（以下，WP基底）の集合 [2] であるとする．また，この推定問題を扱うにあたり，本研究ではベイズ決定理論 [9] の立場からの推定を考える．

なお，問題設定Aにおいては，2乗誤差損失のもとでのベイズ最適な推定に関して既に多くの研究が行われている [6], [7]．一方，問題設定Bにおいてはモデル選択基準を用いて基底を基底の集合の中から一つに決

† 日本電信電話株式会社，横須賀市
Nippon Telegraph and Telephone Corporation, 1-1 Hikarinooka, Yokosuka-shi, 239–0847 Japan
†† 早稲田大学理工学部経営システム工学科，東京都
School of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169–8555 Japan
a) E-mail: kitahara@nttvdt.hil.ntt.co.jp

(注1)：以下では正規直交性を特に強調する場合を除いて単に“基底”と呼ぶ．

定して未知信号の推定を行うという方針が採用されていた [3], [4], [8]. なお，ベイズ的手法を用いた従来研究としては文献 [3], [4] の研究がある．この従来研究で提案された方式では，決定された基底のもとで従来研究 [6], [7] と同様の方法（変換係数のみを未知とした場合のベイズ最適な推定法）で未知信号の推定を行う．この従来法においても WP 基底の集合を仮定しており，その性質を利用して推定信号が効率的に求まる．しかし，問題設定 B では基底と変換係数が未知であることを考慮すると，この従来研究で提案されている推定法は 2 乗誤差損失のもとでベイズ最適ではない．

本論文では問題設定 B においてのベイズ最適な推定に関する基礎検討を目的とし，種々の調整によって様々な応用に耐え得る基礎的な枠組みを提示する．まず **3.** において問題設定 B においてベイズ最適な推定量を提案する．この推定量はそれぞれの基底を用いた場合の推定信号の重み付平均になり，基底を一つに決定する従来の方針とは異なった方針が得られる．しかし，推定を通常の計算法で行うと計算量が膨大になってしまうという問題がある．そこで，事前モデルに対して一定の制約をおくことにより，WP 基底の階層構造を利用して効率的にベイズ最適な推定信号を計算するアルゴリズムを **4.** において提案する．このアルゴリズムによれば従来法 [3], [4] と同等のオーダ $O(N\log_2 N)$ の積和演算量で推定信号を計算できる．**5.** においては，従来法 [3], [4] との理論的側面からの比較を通して本手法の有効性を検討する．更に，現実の問題に本手法を適用する場合に必要となる，種々の調整に関する検討を **5.** で行う．最後に，**6.** において数値実験を通してベイズ最適性以外の推定量の一般的な性質に関する考察を行う．

2. ウェーブレットパケット基底

本研究では WP 基底を以下で説明するような基底の有限集合に含まれる基底の総称として定義する．また，本研究ではデータ長が Nの信号に対して変換を行う場合，変換される信号が周期 Nであると仮定して変換を行う．したがって，以下の説明はこれを前提としてたものになっていることに注意されたい．また，$\boldsymbol{Z}$ は整数の集合を表すとし，$h_0(n)$, $h_1(n)$, $n \in \boldsymbol{Z}$はそれぞれある直交ミラーフィルタ（以下，QMF）[2] の低域，高域フィルタのインパルス応答であるとする．

ここで，$\dot{\delta}_N(n) = \sum_{k\in\boldsymbol{Z}} \delta(n-kN)$ とし，

$$w_{0,0}(n) = \dot{\delta}_N(n), \tag{2}$$

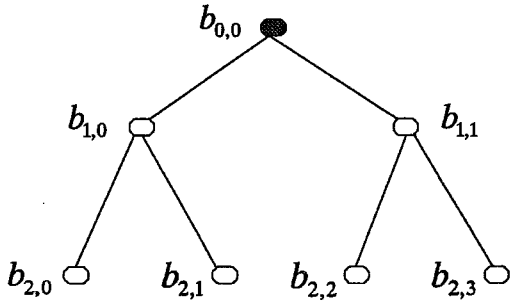

図 1　深さ Jが 2 の WP 木
Fig. 1　WP tree of depth $J = 2$.

とおく．ただし，$\delta(n)$ は単位インパルス信号である．そして，以下の二つの式で与えられる離散時間信号 $w_{j,k}(n)$, $j \in \{0,\cdots,J\}$（ただし，$J \leqq \log_2 N$），$k \in \{0,\cdots,2^j-1\}$ を考える．

$$\begin{aligned} w_{j+1,2k}(n) &= \sum_{l\in\boldsymbol{Z}} h_0(l) w_{j,k}(n-2^j l) \\ w_{j+1,2k+1}(n) &= \sum_{l\in\boldsymbol{Z}} h_1(l) w_{j,k}(n-2^j l) \end{aligned} \tag{3}$$

ここで$\boldsymbol{w}_{j,k,l} = (w_{j,k}(-2^j l), w_{j,k}(-2^j l+1), \cdots, w_{j,k}(-2^j l+N-1))^T$とおき，$b_{j,k} = \{\boldsymbol{w}_{j,k,l}, l \in \{0,1,\cdots,(N/2^j)-1\}\}$ とすると，$b_{j,k}$, $j \in \{0,\cdots,J\}$, $k \in \{0,\cdots,2^j-1\}$ はそれぞれ$\boldsymbol{R}^N$の部分空間の正規直交基底を構成する．更に，正規直交基底 $b_{j,k}$が張る部分空間を $V_{j,k}$とおき，$\oplus$は直和を表すとすると

$$V_{j,k} = V_{j+1,2k} \oplus V_{j+1,2k+1}, \tag{4}$$

という階層的な関係が成立し，これは 2 進木の形で表現できる（図 1）．以下ではこの 2 進木を WP 木と呼ぶ．また，WP 木の深さ jにおいて左から k番目のノードを (j,k)，WP 木の全ノードのインデックスの集合を Sとする．

一つの WP 基底は WP 木のある部分木（これを WP 基底木と呼ぶ）の葉ノードに対応した $b_{j,k}$の和集合として与えることができる(注2)．ここで，与えられた深さの WP 木から構成できる WP 基底の数を Kと表記し，それぞれの WP 基底にインデックスを対応させる．そして，インデックス i, $i \in \{0,1,\cdots,K-1\}$ に対応した WP 基底木の葉ノードのインデックスの集合を S^i_{leaf}と

(注 2)：WP 基底は$\boldsymbol{R}^N$の基底であるから，WP 基底木の根ノードは $(0,0)$ である．

すると，対応するWP基底B_iは$B_i = \cup_{(j,k)\in S^i_{leaf}} b_{j,k}$と表すことができる．

以下では，WP基底の集合はあらかじめ定められた深さのWP木によって構成できるすべてのWP基底を要素とし，$L_0 = \{B_0, \cdots, B_{K-1}\}$と表記する．また，$B_i$の基底ベクトルを行ベクトルとした$N \times N$直交行列を$\boldsymbol{W}_i$とし，$\boldsymbol{x} = (x(0), x(1), \cdots, x(N-1))^T$などのように時間領域の信号を表す$N$次元ベクトルを小文字で表す．そして，WP基底$B_i$に関する変換係数ベクトルはノーテーションを大文字に変えて添字としてiを付ける．つまり，例えば

$$\boldsymbol{X}_i = \boldsymbol{W}_i \boldsymbol{x}, \tag{5}$$

などのように表す．

なお，実際のWP変換においては式(5)の行列演算は必要なく，それぞれのWP基底での変換係数ベクトルを一度に$O(N \log_2 N)$の積和演算量で求めるアルゴリズムが提案されている[10]．

3. ベイズ決定理論に基づく定式化と最適な推定量の提案

3.1 問題設定

式(1)で与えられる離散時間信号ベクトル$\boldsymbol{y}$が観測される．ただし，雑音信号の分散σ^2は既知であるとする．本研究では$(B_i, \boldsymbol{X}_i)$，$B_i \in L_0$，$\boldsymbol{X}_i \in \boldsymbol{R}^N$を未知パラメータとして$\boldsymbol{x}$を推定する問題を扱う．なお，未知信号と推定信号の間の2乗誤差をできるだけ小さくすることを目指した推定を考えていく．

3.2 ベイズリスクの定義

[定義 3.1] $\boldsymbol{x}$を推定量$\hat{\boldsymbol{x}}$(注3)によって推定したときの損失関数$L(\boldsymbol{x}, \hat{\boldsymbol{x}})$を次式で定義する．

$$L(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 \tag{6}$$

□

[定義 3.2] リスク関数を次式で定義する．

$$R(B_i, \boldsymbol{X}_i, \hat{\boldsymbol{x}}) = \int_{\boldsymbol{R}^N} L(\boldsymbol{W}_i^T \boldsymbol{X}_i, \hat{\boldsymbol{x}}) f(\boldsymbol{Y}_i | \boldsymbol{X}_i, B_i) d\boldsymbol{Y}_i \tag{7}$$

ただし，$f(\boldsymbol{Y}_i|\boldsymbol{X}_i, B_i)$は$\boldsymbol{X}_i$が与えられたもとでの$\boldsymbol{Y}_i$の同時確率密度関数を表す． □

なお，一般に未知パラメータ$(B_i, \boldsymbol{X}_i)$，$B_i \in L_0$，$\boldsymbol{X}_i \in \boldsymbol{R}^N$に関して一様に$R(B_i, \boldsymbol{X}_i, \hat{\boldsymbol{x}})$を最小化する推定量は存在しない[14]．したがって，ベイズ基準では未知パラメータの事前分布に関するリスク関数の期待値であるベイズリスク関数を最小化することを考える[9]．ここで，$(B_i, \boldsymbol{X}_i)$の事前同時確率密度関数を$g(\boldsymbol{X}_i, B_i) = g(\boldsymbol{X}_i|B_i) \times P(B_i)$と表す．ただし，$g(\boldsymbol{X}_i|B_i)$は$B_i \in L_0$における未知信号の変換係数ベクトルの事前同時確率密度関数，$P(B_i)$はWP基底B_iの事前確率である(注4)．このように，ベイズ基準を適用する場合は確率分布$P(B_i)$，$i \in \{0, 1, \cdots, K-1\}$により，"真の基底"$B_{i*} \in L_0$が一つ定まり，更に$g(\boldsymbol{X}_{i*}|B_{i*})$からの実現値による"真の変換係数ベクトル"$\boldsymbol{X}^*_{i*}$が定まり，$(B_{i*}, \boldsymbol{X}^*_{i*})$によって未知信号が定まると仮定することになる．

ところで，**1.**で述べた問題設定Aでは未知パラメータをなす基底があらかじめ一つに定められており（これは真の基底B_{i*}を既知としていると解釈できるため，定められた基底をB_{i*}と表す），真の変換係数ベクトル$\boldsymbol{X}^*_{i*}$のみが未知である．したがい，この場合のベイズリスク関数$BR(g, i*, \hat{\boldsymbol{x}})$はリスク関数$R(B_i, \boldsymbol{X}_i, \hat{\boldsymbol{x}})$の$g(\boldsymbol{X}_i|B_i)$に関する期待値であり，次式で与えられる．

$$BR(g, i*, \hat{\boldsymbol{x}}) = \int_{\boldsymbol{R}^N} \int_{\boldsymbol{R}^N} L(\boldsymbol{W}^T_{i*} \boldsymbol{X}_{i*}, \hat{\boldsymbol{x}}) \times f(\boldsymbol{Y}_{i*}|\boldsymbol{X}_{i*}, B_{i*}) d\boldsymbol{Y}_{i*} g(\boldsymbol{X}_{i*}|B_{i*}) d\boldsymbol{X}_{i*} \tag{8}$$

一方，本研究で扱う問題設定Bにおいては真の基底と変換係数ベクトル$(B_{i*}, \boldsymbol{X}^*_{i*})$が未知である．したがって，$L_0$に含まれるそれぞれのWP基底のもとでの問題設定Aにおいてのベイズリスク関数$BR(g, i, \hat{\boldsymbol{x}})$を，更にWP基底の事前分布$P(B_i)$，$i \in \{0, 1, \cdots, K-1\}$によって平均化した期待値が問題設定Bにおいてのベイズリスク関数になる．すなわち，次のように定義される．

[定義 3.3] ベイズリスク関数$BR(g, P, \hat{\boldsymbol{x}})$を次式で定義する．

$$BR(g, P, \hat{\boldsymbol{x}}) = \sum_{B_i \in L_0} \int_{\boldsymbol{R}^N} \int_{\boldsymbol{R}^N} L(\boldsymbol{W}_i^T \boldsymbol{X}_i, \hat{\boldsymbol{x}}) f(\boldsymbol{Y}_i|\boldsymbol{X}_i, B_i) d\boldsymbol{Y}_i \times g(\boldsymbol{X}_i|B_i) d\boldsymbol{X}_i P(B_i) \tag{9}$$

(注3)：本研究で扱う推定量はいずれも$\boldsymbol{y}$の関数だが，表記の簡略化のためにノーテーションには$\boldsymbol{y}$を含めないことに注意されたい．

(注4)：$g(\boldsymbol{X}_i|B_i)$，$P(B_i)$，$i \in \{0, 1, \cdots, K-1\}$がどのようなモデルとして設定されるかは，応用される問題に応じて様々な形をとると考えられる．これらの調整は**5.2**にて述べる従来の応用化手法によって実現可能であるため，本論文では事前モデルは既に与えられているものとして議論を進める．

ただし，$\hat{\boldsymbol{X}}_i$は$\boldsymbol{x}$の推定量$\hat{\boldsymbol{x}}$のB_iに関する変換係数ベクトルである． □

3.3 ベイズ最適な推定量

まず，問題設定Aにおいて2乗誤差損失のもとでベイズ最適(注5)（$BR(g, i*, \hat{\boldsymbol{x}})$を最小化する）な推定量を次式に示す[6], [7].

$$\hat{\boldsymbol{x}}_{i*} = \boldsymbol{W}_{i*}^T \int_{\boldsymbol{R}^N} g(\boldsymbol{X}_{i*}|\boldsymbol{Y}_{i*}, B_{i*}) \boldsymbol{X}_{i*} d\boldsymbol{X}_{i*} \quad (10)$$

ただし，

$$g(\boldsymbol{X}_i|\boldsymbol{Y}_i, B_i) = \frac{f(\boldsymbol{Y}_i|\boldsymbol{X}_i, B_i) g(\boldsymbol{X}_i|B_i)}{\int_{\boldsymbol{R}^N} f(\boldsymbol{Y}_i|\boldsymbol{X}_i, B_i) g(\boldsymbol{X}_i|B_i) d\boldsymbol{X}_i}, \quad (11)$$

は$\boldsymbol{X}_i$の事後同時確率密度関数である．

このように，$\hat{\boldsymbol{x}}_{i*}$は$\boldsymbol{X}_{i*}$の事後分布の平均ベクトル（事後平均ベクトル）の時間領域への逆変換になる．

一方，次の補題において本研究で提案する推定量を示す．この推定量は問題設定Bにおいて2乗誤差損失のもとでベイズ最適であり，$BR(g, i, \hat{\boldsymbol{x}})$をWP基底の事前分布によって平均化した期待値である$BR(g, P, \hat{\boldsymbol{x}})$を最小化する．

[補題 3.1] 式(9)で与えられるベイズリスク関数$BR(g, P, \hat{\boldsymbol{x}})$を最小化する推定量$\hat{\boldsymbol{x}}_{BW}$は次式で与えられる[5].

$$\hat{\boldsymbol{x}}_{BW} = \sum_{B_i \in L_0} \boldsymbol{W}_i^T \int_{\boldsymbol{R}^N} g(\boldsymbol{X}_i|\boldsymbol{Y}_i, B_i) \boldsymbol{X}_i d\boldsymbol{X}_i \times P(B_i|\boldsymbol{Y}_i) \quad (12)$$

ただし，式(11)の分母を$f(\boldsymbol{Y}_i|B_i)$とし，

$$P(B_i|\boldsymbol{Y}_i) = \frac{f(\boldsymbol{Y}_i|B_i) P(B_i)}{\sum_{B_i \in L_0} f(\boldsymbol{Y}_i|B_i) P(B_i)}, \quad (13)$$

はB_iの事後確率である．

（証明）付録参照． □

このように，$\hat{\boldsymbol{x}}_{BW}$はWP基底の事後確率によって$\hat{\boldsymbol{x}}_i$の重み付き平均をとった信号になる．

しかし，WP基底はWP木の深さを増すごとに急激にその数が増え，WP木の深さを最大化した場合のWP基底の数は$O(2^N)$であることが知られている[2]. そして，$\hat{\boldsymbol{x}}_{BW}$の計算においては$L_0$に含まれる一つひとつのWP基底の事後確率を求め，$\hat{\boldsymbol{x}}_i$の重み付けをWP基底の数だけ行う必要がある．したがって，$\hat{\boldsymbol{x}}_{BW}$の計算量は膨大になる．そこで，本研究では式(4)のWP基底の階層構造を活用し，ベイズ最適性を保存しつつ推定信号を効率的に求めるアルゴリズムを次章において提案する．

4. 推定信号の効率的な計算法の提案

4.1 事前モデルに関する制約

提案アルゴリズムにおいて式(4)で表されるWP基底の階層構造を活用するためには未知信号の変換係数，WP基底の事前分布に一定の制約をおく必要があり，これを以下に示す．なお，提案アルゴリズムが適用される上でのQMFの条件は，式(4)の階層構造の構成のみであり，タップ長等の性質は任意である．ここで，$b_{j,k}$，$(j,k) \in S$の基底成分に対応した$\boldsymbol{x}$，$\boldsymbol{y}$の変換係数による$2^{-j}N$次元ベクトルを$\boldsymbol{X}_{j,k}$，$\boldsymbol{Y}_{j,k}$と表す．

まず，未知信号の変換係数の事前分布の制約について述べる．$\boldsymbol{X}_i$，$i \in \{0, \cdots, K-1\}$の事前分布の設定の際，$\boldsymbol{X}_{j,k}$，$(j,k) \in S_{leaf}^i$の事前独立を仮定する．すなわち，$\boldsymbol{X}_{j,k}$の事前同時確率密度関数$g(\boldsymbol{X}_{j,k}|b_{j,k})$，$(j,k) \in S$とすると，$g(\boldsymbol{X}_i|B_i)$，$i \in \{0, \cdots, K-1\}$は次式で与えられる．

$$g(\boldsymbol{X}_i|B_i) = \prod_{(j,k) \in S_{leaf}^i} g(\boldsymbol{X}_{j,k}|b_{j,k}) \quad (14)$$

更に，WP基底の事前分布はツリーモデル[12]と呼ばれる，以下で示すようなパラメトリックな確率分布であるとする．まず，以下を満たすようなWP木の各々のノード$(j,k) \in S$に対応したパラメータ$u_{j,k}$を考える．

$$\begin{cases} u_{j,k} = 1 & j = J \\ u_{j,k} \in (0,1) & \text{otherwise} \end{cases} \quad (15)$$

$u_{j,k}$を用いてWP基底$B_i \in L_0$の事前確率は次式で与えられる．

$$P(B_i) = \prod_{(j,k) \in S_{leaf}^i} u_{j,k} \prod_{(l,m) \in S_{inter}^i} (1 - u_{l,m}) \quad (16)$$

ただし，S_{inter}^iはB_iのWP基底木の内節ノードのインデックスの集合である．

なお，5.において本研究と同じ問題設定を扱った従来法[3], [4]と事前モデルの制約に関して比較を行う．

(注5)：ベイズ最適とは，ベイズリスク関数を最小化するということである．

4.2 提案アルゴリズム

ここで，提案アルゴリズムの記述，及びその最適性の証明において必要である，新たなノーテーションについて述べる．$2^{-j}N$次元ベクトル$\boldsymbol{D} = (D(0), \cdots, D(2^{-j}N-1))^T \in \boldsymbol{R}^{2^{-j}N}$に次式のようにQMFを用いてフィルタリングを行って得られる信号，$D^{\uparrow i}(n)$，$i=0,1$，$n \in \boldsymbol{Z}$を考える．

$$D(n)^{\uparrow i} = \sum_{l \in \boldsymbol{Z}} h_i(2l-n)\dot{D}(l), \quad i=0,1 \tag{17}$$

ただし，$\dot{D}(n)$は$D(n)$に周期化拡張（周期$2^{-j}N$）を行った信号であり，$\dot{D}(n+2^{-j}Nl) = D(n)$，$l \in \boldsymbol{Z}$を満たす．このようにして得られる，$D^{\uparrow i}(n)$，$i=0,1$，$n=0,1,\cdots,2^{-j+1}N$を$n$番目の要素とする$2^{-j+1}N$次元ベクトルは，フィルタリングされた信号ベクトルのノーテーションの右上に$\uparrow i$を付け，それぞれ$\boldsymbol{D}^{\uparrow i}$，$i=0,1$などと表す(注6)．

-the Proposed Algorithm-

- **step**(0)

観測信号$\boldsymbol{y}$を得ることにより，WP木の全ノードのインデックス$(j,k) \in S$について以下を計算する．

$$\tilde{\boldsymbol{X}}_{j,k} = \int_{\boldsymbol{R}^{2^{-j}N}} g(\boldsymbol{X}_{j,k}|\boldsymbol{Y}_{j,k}, b_{j,k})\boldsymbol{X}_{j,k}d\boldsymbol{X}_{j,k}, \tag{18}$$

$$f(\boldsymbol{Y}_{j,k}|b_{j,k}) = \int_{\boldsymbol{R}^{2^{-j}N}} f(\boldsymbol{Y}_{j,k}|\boldsymbol{X}_{j,k}, b_{j,k})g(\boldsymbol{X}_{j,k}|b_{j,k})d\boldsymbol{X}_{j,k} \tag{19}$$

ただし，$g(\boldsymbol{X}_{j,k}|\boldsymbol{Y}_{j,k}, b_{j,k})$は$\boldsymbol{X}_{j,k}$の事後同時確率密度関数，$f(\boldsymbol{Y}_{j,k}|\boldsymbol{X}_{j,k}, b_{j,k})$は$\boldsymbol{X}_{j,k}$が与えられたもとでの$\boldsymbol{Y}_{j,k}$の同時確率密度関数である．

- **step**(1)

WP木の最も深いノードから根ノードに向かって，ノードごとに以下の再帰的な計算を行う．

– step(1.0)

WP木のノード(J,k)，$k \in \{0,\cdots,2^J-1\}$のそれぞれに対応した$P_{J,k}$，$\tilde{\boldsymbol{X}}_{J,k}$を次式によって求める．

$$P_{J,k} = u_{J,k}f(\boldsymbol{Y}_{J,k}|b_{J,k}) \tag{20}$$

$$\tilde{\boldsymbol{X}}_{J,k} = P_{J,k}\bar{\boldsymbol{X}}_{J,k} \tag{21}$$

– step(1.k)，$1 \leqq k \leqq J$

$j = J-k$とおく．WP木のノード(j,k)，$k \in \{0,\cdots,2^j-1\}$のそれぞれに対応した$P_{j,k}$，$\tilde{\boldsymbol{X}}_{j,k}$を次式によって求める．

$$P_{j,k} = u_{j,k}f(\boldsymbol{Y}_{j,k}|b_{j,k}) + (1-u_{j,k})P_{j+1,2k}P_{j+1,2k+1} \tag{22}$$

$$\tilde{\boldsymbol{X}}_{j,k} = u_{j,k}f(\boldsymbol{Y}_{j,k}|b_{j,k})\bar{\boldsymbol{X}}_{j,k} + (1-u_{j,k})(P_{j+1,2k+1}\tilde{\boldsymbol{X}}^{\uparrow 0}_{j+1,2k} + P_{j+1,2k}\tilde{\boldsymbol{X}}^{\uparrow 1}_{j+1,2k+1}) \tag{23}$$

- **step**(2)

$$\hat{\boldsymbol{x}}_{BW} = \tilde{\boldsymbol{X}}_{0,0}/P_{0,0} \tag{24}$$

以上で示した提案アルゴリズムにおいてはベイズ最適性が保証されており，これは以下の定理で示される．

[定理 4.1]　提案アルゴリズムはベイズリスク関数$BR(g,P,\hat{\boldsymbol{x}},L_0)$を最小化する．

（証明）付録参照．　□

変換係数の事前分布に仮定された制約により，**step**(0)を行うことで$\boldsymbol{W}_i\hat{\boldsymbol{x}}_i$，$i \in \{0,1,\cdots,K-1\}$（式(10)を参照）が求まる．そして，**step**(1)以降において$\hat{\boldsymbol{x}}_i$，$i \in \{0,1,\cdots,K-1\}$がWP基底の事後分布によって平均化される．提案アルゴリズムによってもたらされる効率化は，個々のWP基底の事後確率を求めて$\hat{\boldsymbol{x}}_i$の重み付き平均を通常の計算で求める必要がないところによる．具体的には，**step**(1)においてWP木の最も深いノードから根ノードに向かって，ノードごとに式(20)，(21)，(22)，(23)の再帰的な計算を行えばよい．この効率的な計算法はWP基底の事前分布に式(16)のツリーモデルを仮定したことにより，式(4)で表現されるWP基底の階層構造が活用されることによって可能になる(注7)．なお，計算量に関しては**5.**において述べる．

5. 考　　察

5.1 理論的側面から従来法との比較

ここでは本研究と同様の問題設定を扱い，ベイズ的手法を適用した従来研究[3], [4]について簡単に述べる．そして，推定量の最適性，事前モデルに関する制約，計算量の観点から本研究と従来法の理論的相違を明確化し，本研究の提案の有効性について検討する．

(注6)：$X^{\uparrow 0}_{j,2k}$，$X^{\uparrow 1}_{j,2k+1}$はそれぞれ$b_{j-1,k}$の変換領域における$X_{j,2k}$，$X_{j,2k+1}$の表現であることに注意する．

(注7)：提案アルゴリズムにおけるWP基底の階層構造の活用に関する詳細については定理4.1の証明を参照されたい．

従来法 [3], [4] ではベイズ的手法を適用しているものの，**3.** で示したようなベイズ決定理論に基づいた明確な定式化が行われていない．なお，次に従来の推定法を示し，ベイズ決定理論からの従来法の解釈を示す．

観測信号が与えられると，まず次式によって WP 基底を一つに決定する（$g(\boldsymbol{X}_i|B_i)$, $i \in \{0,1,\cdots,K-1\}$ は与えられているとされる）．

$$B_{\hat{i}*} = \underset{B_i \in L_0}{\arg\max} \log f(\boldsymbol{Y}_i|B_i) \tag{25}$$

そして，従来法の未知信号の推定量を$\hat{\boldsymbol{x}}_{BSP}$と表すと，$\hat{\boldsymbol{x}}_{BSP} = \hat{\boldsymbol{x}}_{\hat{i}*}$（式 (10) を参照）である．

ここで，式 (25) による基底選択は WP 基底の事前分布を一様分布として事後確率最大の WP 基底を選択することと等価である．したがって，$B_{\hat{i}*}$は WP 基底の事前分布を一様分布とし，0-1 損失を損失関数とした場合の真の基底 B_{i*}のベイズ最適な推定量になっている [9]．しかし，未知信号と推定信号の間の 2 乗誤差を損失関数としたもとでのベイズ決定理論による定式化によれば，従来法のように基底を一つに決定することは最適でない．そして，2 乗誤差損失のもとでベイズ最適である$\hat{\boldsymbol{x}}_{BW}$によれば，基底をその事後確率で重み付けすることが最適であることがわかる．

ここで，信号ベクトルの次元 Nと$\hat{\boldsymbol{x}}_{BW}$，$\hat{\boldsymbol{x}}_{BSP}$の推定精度の関係について簡単に述べておく．観測データの数が増えると，ベイズ最適な推定量による推定精度が向上する傾向があることはベイズ決定理論においてはよく知られている．また，本研究においての問題設定のように事後確率による重み付けを推定に用いることがベイズ最適であるとき，ベイズ最適な推定量と事後確率最大化による推定量の推定精度の差が小さくなる傾向があることも知られている．そして，いずれの傾向も本手法においても見受けられることが次章で提示する数値実験で確認できる．一方，主に統計的推定・予測と情報源符号化の分野において，ベイズ最適な決定による損失，リスク関数，ベイズリスク関数の，観測データ数に関する理論的な漸近評価を行った研究が従来から盛んに行われている [15]〜[17]．これらの研究では，ある条件(注 8)のもとではベイズ最適な決定と事後確率最大化による決定が漸近的には等価になることを証明している場合がある．これらの結果はそれぞれ限られた条件下において証明され，一般的には非常に難解な問題であることが知られている．同様に，本研究で扱っている問題設定において理論的な漸近評価は困難であるが，一般には従来法が漸近的にベイズ最適である保証はない．

次に事前モデルに関する制約の観点から従来法と比較する．なお，従来法においても推定信号を効率的に求めるため，変換係数の事前分布に **4.1** で述べた制約をおいており，本研究と同様の制約をおいている．また，従来法においては WP 基底の事前分布は一様分布としていると解釈できることは既に述べた．**4.1** で述べたツリーモデル [12] は一様分布をはじめとして，様々な確率分布を表現できる．ツリーモデルはユニバーサル無ひずみ情報源符号化の一方式であるベイズ符号化アルゴリズムのために松嶋らに提案された [12]．このモデルは確率分布の表現において高い自由度をもっていることが知られ，木構造モデルを用いた様々なベイズ推定に応用されている [16]．

最後に計算量の観点から$\hat{\boldsymbol{x}}_{BW}$と$\hat{\boldsymbol{x}}_{BSP}$を比較する．ここでは$\boldsymbol{Y}_i$, $i \in \{0,1,\cdots,K-1\}$がすべて求まった後の計算量を積和演算量で評価する．なお，ここでの計算量の評価においては変換係数の事前分布に関しては従来法と同じモデル（Bernoulli-Gaussian モデル）を用いた場合を考える．このモデルをはじめとして，変換係数の事前分布に自然共役事前分布 [9] を仮定すると式 (18) で生じる積分計算は単純な算術計算になる [3], [4], [6], [9]．すなわち，下記の結果は従来研究で用いられた様々なモデルのもとにおいても同様である．

従来法の推定信号は “最適基底アルゴリズム” [2] を用いて $O(N\log_2 N)$ で求まる．次に提案法の計算量を述べる．$\boldsymbol{X}_{j,k}$は $2^{-j}N$次元ベクトルであるから，**step**(0) での計算量は $O(N\log_2 N)$ である．提案アルゴリズムによる効率化は，個々の WP 基底の事後確率による$\hat{\boldsymbol{x}}_i$の重み付き平均を通常の計算で求めるのではなく，**step**(1) において WP 木の一番深いノードから根ノードに向かって式 (20), (21), (22), (23) で与えられる再帰的な計算をノードごとに行えばよいことによる．この効率的な計算法は式 (4) で表現される WP 基底の階層構造を活用している．なお，**step**(1) の再帰計算においての計算量も $O(N\log_2 N)$ である（これは QMF によるフィルタリングの計算量も含む）．これはノード (j,k) においての計算量が $O(2^{-j}N)$ であり，WP 木の深さ jにおいてのノードの数は 2^j，WP 木の深さはたかだか $\log_2 N$であることによる．したがい，提案法の計算量は全体で $O(N\log_2 N)$ で抑えられる．

(注 8)：観測信号に対する，二つの異なる確率モデルが同じ確率分布を表現できないことを要求するなど，様々な条件が要求される．

すなわち，従来法と同じ制約条件下において，オーダの意味では従来法と同等のオーダの計算量で推定信号を計算できることになる．

以下では一様分布に限らず一般的な事前分布を WP 基底の事前分布として仮定した場合の事後確率最大の基底を B_{i*} とし，$\hat{\boldsymbol{x}}_{BSP} = \hat{\boldsymbol{x}}_{i*}$ とする．

5.2 現実の問題への応用に向けた検討

本手法においては，WP 基底の集合 L_0，WP 基底の事前分布 $P(B_i)$，$i \in \{0, 1, \cdots, K-1\}$，変換係数ベクトルの事前分布 $g(\boldsymbol{X}_i|B_i)$，$B_i \in L_0$，$\boldsymbol{X}_i \in \boldsymbol{R}^N$(注9)，雑音信号モデルの分散パラメータ σ^2 さえ与えられれば，**3.1** で述べた問題設定に当てはまるすべての信号推定問題においてベイズ最適な推定が行える．しかし，ベイズ最適性は事前分布に関する平均的な意味での最適性である．したがって，現実の問題において良好な推定を行うには，事前分布等を問題に応じて適切に設定する必要がある．本研究の目的は **3.1** で述べた問題設定に対するベイズ最適な推定量，及び効率的なアルゴリズムの提案などにより，様々な問題に応用可能な基礎的な枠組みを提示することである．したがって，本手法を個々の問題に用いるための調整は今後の課題としたい．しかし，従来の手法を本手法の枠組みに組み込むことにより，様々な応用手法が考え得るため，以下で簡単に説明する．

本手法の枠組みにおいては，変換係数ベクトルの事前分布は式 (14) の制約を満たす必要がある．しかし，良好な推定を行うには式 (14) の制約が未知信号の性質に関して妥当である必要がある．文献 [3], [4], [6], [7] に代表される多くの従来研究において同様の制約がおかれているが，B_{i*} によって未知信号のエネルギーが少ない変換係数に集中され，変換係数が無相関化されることを要求することによって式 (14) の制約が正当化される．そして，通常は $g(\boldsymbol{X}_{j,k}|b_{j,k})$，$(j,k) \in S$ の各々に対して同じパラメトリックモデルが適用され，モデルのパラメータのみ $(j,k) \in S$ に依存するような形がとられる．なお，具体的なパラメトリックモデルとしては t 分布モデル [7]，Gaussian Mixture モデル [6]，Bernoulli-Gaussian モデル [3], [4] などが $g(\boldsymbol{X}_{j,k}|b_{j,k})$，$(j,k) \in S$ に適用され，未知信号に関する事前情報を反映して変換係数の事前独立と平均パラメータがゼロベクトルであることが仮定される．また，これらの研究では事前モデルの平均パラメータ以外のパラメータの設定には経験的ベイズ法 [9] が用いられている．経験的ベイズ法では，観測データから何らかの方法で事前分布のパラメータが推定値が求められ，この推定値を用いて通常のベイズ推定が行われる．以上の手法の多くは本手法の枠組みに容易に組み込むことが可能である．

次に，WP 基底の事前モデル（式 (28) のツリーモデル）のパラメータの設定法について述べる．前節で述べたとおり，本研究と同様の問題設定を扱った従来法 [3], [4] では WP 基底の事前分布を一様分布としていると解釈できる．ベイズ決定理論では，とり得る値が有限である確率変数に対する一様分布は無情報事前分布と称され，B_{i*} に関して何も事前情報がない場合に用いられる．一様分布を与えるツリーモデルのパラメータは次式のように再帰的に与えられる．

$$\bar{u}_{j,k} = \frac{1}{\frac{1}{\bar{u}^2_{j+1,2k}} + 1} \tag{26}$$

ただし，$\bar{u}_{J,k} = 1$ である．なお，パラメータ $u_{j,k}$ を大きくする（小さく）すると，$b_{j,k}$ を含む WP 基底の事前確率を大きく（小さく）し，同時に，WP 木のノード (j,k) の子孫ノードに対応した基底を含む WP 基底の事前確率を小さく（大きく）する性質がある．したがって，偏りのある事前分布を設定する際はこの性質を考慮し，$\bar{u}_{j,k}$ を基準としてパラメータを調整して所望の事前分布を近似すればよい．

一方，WP 基底の集合 L_0 に関しては，式 (4) の階層構造さえ満たされていれば提案アルゴリズムが適用でき，ベイズ最適な推定が効率的に行える．しかし，事前分布の設定と同様に，良好な推定を行うには適切な QMF を用いる必要がある（例えば，前述の従来研究のように B_{i*} によって未知信号が少ない変換係数で近似されることを要求する場合，これを考慮した QMF を利用する必要がある）．従来は未知信号の性質に関する事前情報に応じて経験的に用いる QMF が決められている傾向がある．一方，候補の QMF を複数用意しておき，これらに事前確率を与えてベイズ最適な推定量を構成することも考え得る．ここでは紙面の都合で省略するが，この場合のベイズ最適な推定量は補題 3.1 と同様に容易に導かれる．

最後に，本手法では雑音信号の分散 σ^2 は既知であるとしたが，現実の問題では雑音信号の分散 σ^2 が正確にわかっていないことが多い．この問題に対するベイズ推定における解決法として，σ^2 に対しても事前分布を

(注9)：提案アルゴリズムを用いるためには **4.1** の制約条件を満たすモデルが与えられる必要があることに注意されたい．

仮定して推定を行う従来手法がある [7]. しかし，この手法ではベイズ最適な推定量が解析的に得られず積分計算が避けられないことが多い．したがって，従来の研究では何らかの推定値$\hat{\sigma}^2$を求め，σ^2が既知であると同様に扱われることが多い．具体的には，雑音信号のみを観測することができる場合は雑音信号から直接的に何らかの推定値を求める．また，未知信号が混合した状態でσ^2を推定する必要がある場合は，現実の問題においては未知信号の高周波数エネルギーは小さい傾向があるため，観測信号の高周波数成分からロバスト推定量を用いて推定値$\hat{\sigma}^2$を求める従来手法がある [3], [4], [6]. 以上の手法は容易に本手法の枠組みに組み込むことが可能である．

以上において，本手法を現実の問題に適用する際に検討すべき問題の各々に関する従来法について述べた．これらの手法を組み合わせることが，本手法を実用化する上での一つの方法であることがわかる．なお，これまでに挙げた従来研究ではB_{i*}によって未知信号が少ない変換係数で近似されることを仮定する．したがって，従来の応用化手法の組合せで本手法を応用する場合，B_{i*}によって未知信号のエネルギーが少ない係数に集中されるときに良好な推定が行える．すなわち，地震探査の際に用いられる地震波データ，医療画像の生成に用いられるNMR（Nuclear Magnetic Resonance）スペクトルデータなど，WP基底のような局所的な基底によって少ない変換係数で効率的に表現されるような信号の推定に有用になる．

6. 数値実験

6.1 数値実験の目的

ここでの数値実験では，ある条件のもとで本手法と従来法 [3], [4] を比較することにより，本手法がもつ一般的な性能と性質を評価することを目的としている．これは，**5.2** で述べたような形で何らかの具体的な現実問題に本手法を適合させ，実データに対して性能評価を行うことも重要であるが，まず様々な現実問題において本手法が共通に有するような性質について評価することが応用上においても重要であると考えられるからである．したがって，以下の数値実験で適用される実験条件は，あくまでも本手法の従来法に対する優位性に関する理論的な結果，あるいは仮説の正当性を示すために利用された一つの条件である．そして，様々な応用場面においても以下の数値実験における傾向が同様に見受けられると考えられることに注意されたい．

6.2 数値例によるベイズ最適性の確認（実験A）

6.2.1 実験内容

$\hat{\boldsymbol{x}}_{BW}$の最も重要な有効性は，この推定量が式 (9) のベイズリスク $BR(g, P, \hat{\boldsymbol{x}})$ を最小化することである．これは，以下の数値例により $BR(g, P, \hat{\boldsymbol{x}}_{BW})$, $BR(g, P, \hat{\boldsymbol{x}}_{BSP})$ をそれぞれ標本値による（算術）平均2乗誤差に置き換え，これらを比較することによって確認できる．

観測信号データは後に示す事前モデルの設定のもとで，次のように生成される．具体的にはWP基底の事前分布に従ってWP基底をl個定め，これらのWP基底のそれぞれにおいての変換係数の事前分布に従って変換係数ベクトルをm個発生させることによって$l \times m$個の未知信号が得られる．更に，これらの未知信号のそれぞれにn個の雑音信号を加算することによって$M = l \times m \times n$個の観測信号データ$\boldsymbol{y}^k$, $k = 0, 1, \cdots, M-1$が得られる．なお，$\boldsymbol{y}^k$を観測したときの推定値ベクトルをそれぞれ$\hat{\boldsymbol{x}}^k_{BW}$, $\hat{\boldsymbol{x}}^k_{BSP}$とし，平均2乗誤差 $BMSE(BW) = 1/(NM)\sum_{k=1}^{M} \|\boldsymbol{x}^k - \hat{\boldsymbol{x}}^k_{BW}\|^2$, $BMSE(BSP) = 1/(NM)\sum_{k=1}^{M} \|\boldsymbol{x}^k - \hat{\boldsymbol{x}}^k_{BSP}\|^2$を比較する．

6.2.2 実験条件

未知信号の変換係数の事前モデルとしては従来法 [3], [4] において用いられていたBernoulli-Gaussian（以下，B-G）モデルというパラメトリックモデルを適用する．それぞれの$\boldsymbol{X}_{j,k}$, $(j,k) \in S$の事前モデルに対して異なるパラメータを設定することなどによって複雑なモデルを仮定することが可能だが，ここでは以下に示す簡単なモデルを用いる．ここで仮定するB-Gモデルでは$\boldsymbol{X}_{i*}$のn番目の要素$X_{i*}(n)$は互いに独立でそれぞれ次式で与えられる．

$$X_{i*}(n) = r(n)q(n), n = 1, \cdots, N \tag{27}$$

ただし，$r(n)$はそれぞれ互いに独立で平均0，分散σ_Xの正規分布に従い，$q(n)$は

$$\Pr\{q(n)\} = \begin{cases} 1-\lambda & q(n) = 0 \\ \lambda & q(n) = 1 \end{cases} \tag{28}$$

を満たすようなベルヌーイ系列である．なお，これらのパラメータはWP基底のインデックス$i \in \{0, 1, \cdots, K-1\}$には依存しないとする．ここでは$\sigma_X = 3$, $\lambda = 0.3$であるようなモデルを適用した．WP木の深さJは$J = 4, 5, \cdots, 9$の場合を適用し，WP基

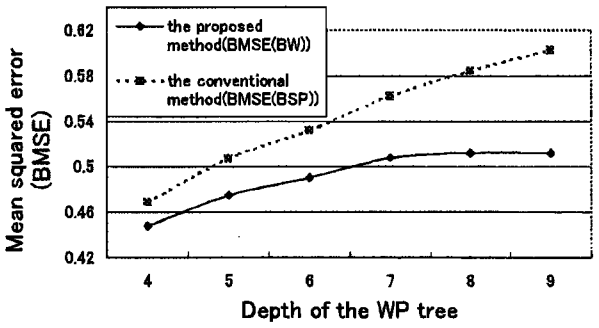

図2 実験Aの結果 ($N = 512, J = 4, 5, \cdots, 9$)
Fig. 2 Simulation results of experiment A ($N = 512, J = 4, 5, \cdots, 9$).

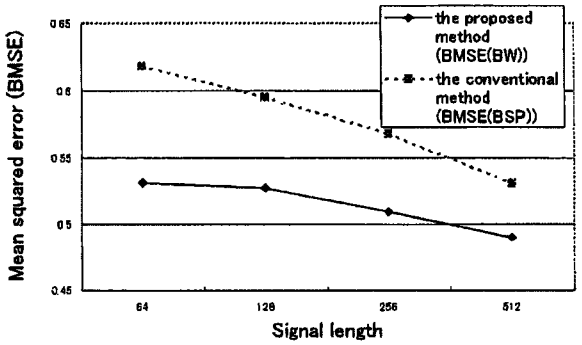

図3 実験Aの結果 ($N = 64, 128, 256, 512, J = 4$)
Fig. 3 Simulation results of experiment A ($N = 64, 128, 256, 512, J = 4$).

底の事前分布はいずれの場合も一様分布を適用した．更に，$\sigma = 1$，$M = 20000$，$N = 64, 128, 256, 512$ とした．QMF は Daubechies の 8 タップフィルタを用いた．

6.2.3 実験結果と考察

図 2，図 3 に実験結果を示す．図 2 は $N = 512$ とした場合の，平均2乗誤差 $BMSE(BW)$，$BMSE(BS)$ と WP 木の深さ Jの関係を表しており，図 3 は $J = 4$ とした場合の $BMSE(BW)$，$BMSE(BS)$ と信号の長さ Nとの関係を表している．すべての条件において $\hat{\boldsymbol{x}}_{BW}$は$\hat{\boldsymbol{x}}_{BSP}$に比べて平均 2 乗誤差を小さくしており，理論どおりの結果が得られている．また，図 2 では WP 木の深さが増すにつれて$\hat{\boldsymbol{x}}_{BW}$と$\hat{\boldsymbol{x}}_{BSP}$の平均 2 乗誤差の差が開く傾向があるが，この傾向に関しては次節の実験においてより詳しい考察を行う．更に，図 3 では Nが大きくなるにつれて $BMSE(BW)$ と $BMSE(BS)$ が互いに小さくなり，同時に互いの差も小さくなる傾向が見られる．しかし，**5.1** で述べたとおり，Nが大きくなるにつれて $BMSE(BS)$ は $BMSE(BW)$ に近づくが，従来法$\hat{\boldsymbol{x}}_{BSP}$と本手法 $\hat{\boldsymbol{x}}_{BW}$が Nに関して漸近的に等価になる保証はない．

6.3 数値実験による推定量の性質の一評価（実験 B）

6.3.1 実験内容

ここでは，$\hat{\boldsymbol{x}}_{BW}$の性能が保証された評価基準 $BR(g, P, \hat{\boldsymbol{x}})$ 以外の評価基準においての$\hat{\boldsymbol{x}}_{BW}$の性能について考察する．具体的には，式 (8) で定義した基準 $BR(g, i*, \hat{\boldsymbol{x}})$ を考える．これは真の基底 B_{i*}が既知であるときのベイズリスク関数であり，WP 基底の事前分布による平均化が行われていない．$BR(g, i*, \hat{\boldsymbol{x}})$ によって推定量を評価することによって次のようなことを知ることができる．定性的には，未知信号の B_{i*}に関する変換係数ベクトルが $g(\boldsymbol{X}_{i*}|B_{i*})$ から発生しやすいようなものであった場合（つまり B_{i*}においての未知信号の変換係数の事前分布の設定が妥当であった場合）の推定量の平均的な性能を知ることができる．

なお，$\hat{\boldsymbol{x}}_{BW}$は WP 基底のある事前分布 $P(B_i), i \in \{0, 1, \cdots, K-1\}$ に関する $BR(g, i, \hat{\boldsymbol{x}})$ の期待値 $BR(g, P, \hat{\boldsymbol{x}})$ を最小化する．すなわち，一つひとつの WP 基底を真の基底として$\hat{\boldsymbol{x}}_{BSP}$と $BR(g, i*, \hat{\boldsymbol{x}})$ について比較した場合も，多くの基底 $B_{i*} \in L_0$について$\hat{\boldsymbol{x}}_{BW}$が勝ると考えられる．そこで，逆に$\hat{\boldsymbol{x}}_{BSP}$が $BR(g, i*, \hat{\boldsymbol{x}})$ のもとで有利になるのはどのような場合であるか知ることにより，$\hat{\boldsymbol{x}}_{BW}$の性質についてより深い理解が得られると考えられる．

まず，以下で述べる$\hat{\boldsymbol{x}}_{BSP}$の性質に着目した．$\hat{\boldsymbol{x}}_{BSP}$ の性質により，観測信号を得て真の基底 B_{i*}の事後確率が他の WP 基底のそれと比較して最大であれば，$\hat{\boldsymbol{x}}_{BSP}$は$\hat{\boldsymbol{x}}_{i*}$と等しい．したがって，$C(i*) \subset \boldsymbol{R}^N$を B_{i*} が事後確率最大になるような$\boldsymbol{Y}_{i*}$の集合としたとき，

$$\begin{aligned}\Pr\{\hat{i*} = i*|B_{i*}\} &= \int_{\boldsymbol{R}^N}\int_{C(i*)} f(\boldsymbol{Y}_{i*}|\boldsymbol{X}_{i*}, B_{i*})d\boldsymbol{Y}_{i*} \\ &\quad \times g(\boldsymbol{X}_{i*}|B_{i*})d\boldsymbol{X}_{i*}, \end{aligned} \tag{29}$$

が十分大きい（B_{i*}が事後確率最大になる確率が大きい）場合にのみ$\hat{\boldsymbol{x}}_{BSP}$が $BR(g, i*, \hat{\boldsymbol{x}})$ をより小さくする傾向があると推測できる．

以上の考察に関して見通しを得るため，以下の方法で実験を行った．B_{i*}をランダムに決め，実験 A と同様に $BR(g, i*, \hat{\boldsymbol{x}}_{BW})$，$BR(g, i*, \hat{\boldsymbol{x}}_{BSP})$ をそれぞれ標本値による平均 2 乗誤差に置き換えた $MSE(BW)$，$MSE(BSP)$ を求め，更に次に説明する *count* を求

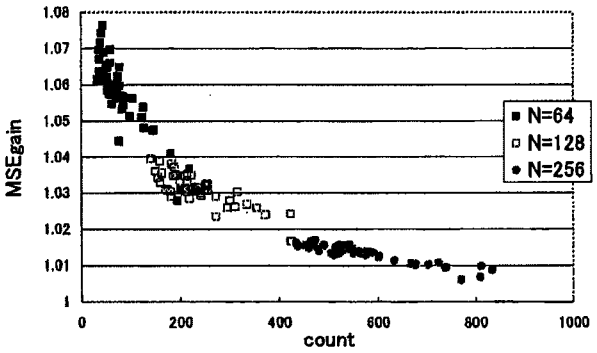

図 4　実験 B の結果 ($N = 64, 128, 256$)
Fig. 4　Simulation results of experiment B ($N = 64, 128, 256$).

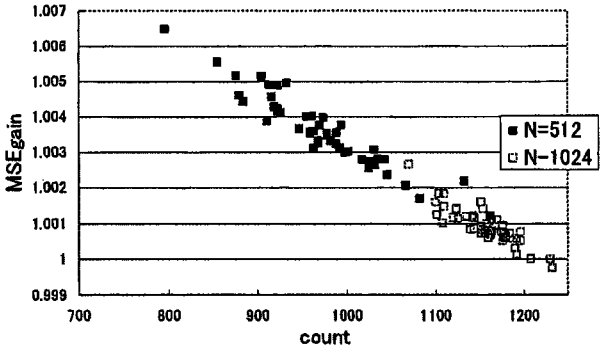

図 5　実験 B の結果 ($N = 512, 1024$)
Fig. 5　Simulation results of experiment B ($N = 512, 1024$).

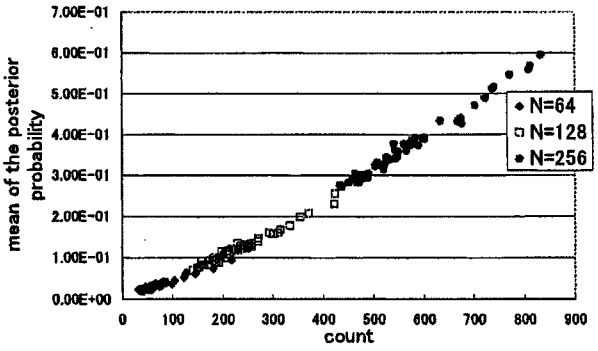

図 6　B_{i*}の事後確率の平均と $count$ のプロット ($N = 64, 128, 256$)
Fig. 6　Plots according to the average of the posterior probability of B_{i*} and $count$ ($N = 64, 128, 256$).

める．ここで $MSE(BW)$，$MSE(BSP)$ を求める際に用いられた観測信号ベクトルの数（標本数）を M とし，$count$ は M 回の観測のうち B_{i*} が事後確率最大になった回数である．上述の考察によれば，$count$ が M に近い場合にのみ $AMSE(BSP)$ が $AMSE(BW)$ より小さい傾向があると考えられる．$AMSE(BW)$，$AMSE(BSP)$，$count$ を L_0 からランダムに選択された 100 個の WP 基底のそれぞれを B_{i*} とした場合について求めて傾向を見る．

6.3.2 実験条件

ここで示す数値例において用いた事前モデル等に関する実験条件を以下に示す(注10)．

$\boldsymbol{X}_{i*}$ の事前モデル，及びそのパラメータ設定，雑音信号の分散，QMF に関しては **5.1** の実験と同様の場合の実験結果を示す．WP 木の深さ j は $J = 6$ の場合を適用し，WP 基底の事前分布はいずれの場合も一様分布を適用した．更に，$M = 1250$ ($l = 50, m = 25$)，$N = 64, 126, 256, 512, 1024$ とした．

6.3.3 実験結果と考察

図 4，図 5 に実験結果を示す．グラフにおける一つのプロット点はある WP 基底を B_{i*} に定めたときの $MSEgain = MSE(BSP)/MSE(BW)$（縦軸），$count$（横軸）をもとにプロットされている．したがい，$MSEgain$ が 1 より大きいときに $\hat{\boldsymbol{x}}_{BW}$ が平均 2 乗誤差をより小さくしたことになる．プロット点の図形によってプロットの集団が分けられているが，これは信号の長さ N による分類に対応し，対応関係は図に記載されたとおりである．

実験結果によれば，推測どおりの結果が得られている．$MSE(BSP)$ が $MSE(BW)$ より小さい傾向があるのは，$count$ が $M = 1250$ に非常に近く，B_{i*} が事後確率最大になりやすい場合のみである．しかし，$MSE(BSP)$ と $MSE(BW)$ の差は比較的小さい傾向が見受けられる．また，$count$ が小さいほど $\hat{\boldsymbol{x}}_{BW}$ の方が推定精度がよく，$MSE(BW)$ と $MSE(BSP)$ の差が大きくなる傾向が見受けられる．また，WP 木が深いほど，及び N が小さいほど $count$ が少ない傾向があるため，$MSE(BW)$ と $MSE(BSP)$ の差が大きくなり，**6.1** の実験結果に対応する．

以上の傾向に関する考察のため，図 6，図 7 に $count$（横軸）と $M = 1250$ 回の観測においての B_{i*} の事後確率の平均（縦軸）を軸にプロットした図を示す．ここでも図 4，図 5 と同様にプロットの集団が N によって分類されている．図 6，図 7 によれば，$count$ が少

(注 10)：他の実験条件を適用した場合の実験も行ったが，ここで示す結果と同様の傾向が見受けられたことから以下の実験条件においての実験結果のみを示す．

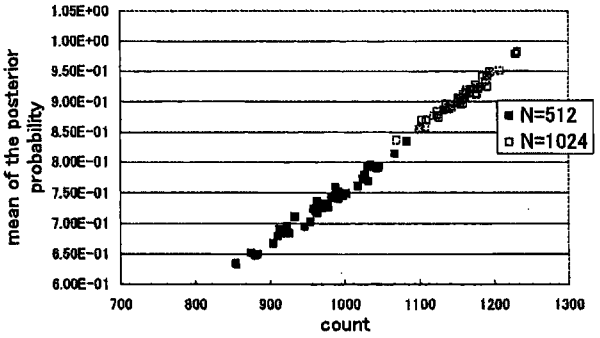

図 7 B_{i*}の事後確率の平均と *count* のプロット ($N = 512, 1024$)

Fig. 7 Plots according to the average of the posterior probability of B_{i*} and *count* ($N = 512, 1024$).

ないほど B_{i*}以外の WP 基底の事後確率が大きい傾向がある．この結果は，*count* が少ないほど WP 基底の事後確率による重み付けを行う効果が顕著になる傾向を示している．一方，図 6，図 7 で示されているように，本研究で扱ったいずれの実験条件においても，*count* が $M = 1250$ に非常に近い場合は B_{i*}の事後確率の平均が 1 に近かった．この結果は *count* が $M = 1250$ に非常に近い場合は推定に事後確率最大の WP 基底のみを用いても，すべての WP 基底を用いても大差はないという傾向を示す重要な結果である．

すなわち，$\hat{\boldsymbol{x}}_{BSP}$にとって最も有利な条件下でも $\hat{\boldsymbol{x}}_{BSP}$は$\hat{\boldsymbol{x}}_{BW}$に比べてわずかに良い性能を示すのみであり，ほとんどの場合において$\hat{\boldsymbol{x}}_{BW}$がより良好な推定が行えると考えられる．更に，信号の長さ Nが小さいほど，及び WP 基底の深さ Jが大きいほど$\hat{\boldsymbol{x}}_{BW}$と$\hat{\boldsymbol{x}}_{BSP}$の性能の差が顕著である．以上をまとめると，あらゆる状況において$\hat{\boldsymbol{x}}_{BW}$の方が安定した推定が行え，$\hat{\boldsymbol{x}}_{BW}$はロバストであると考えられる．

7. むすび

本研究で扱った問題設定に関し，従来においてはベイズ最適な推定に関して考察されていなかった．そこで，本研究ではまず 2 乗誤差損失に関するベイズ最適な推定量$\hat{\boldsymbol{x}}_{BW}$を求めた．しかし，この推定量には計算量の面で問題があるため，WP 基底の性質を活用して推定信号を効率的に計算するアルゴリズムを提案した．このアルゴリズムによれば，従来法と同じ制約条件下で同等のオーダの積和演算量でベイズ最適性が保証された推定信号を計算できる．また，ベイズ最適性以外の$\hat{\boldsymbol{x}}_{BW}$の性質に関する考察を数値実験を通して行った．実験結果によれば，様々な状況において$\hat{\boldsymbol{x}}_{BW}$は従来法と比較して安定した推定を行える傾向が見受けられた．なお，$\hat{\boldsymbol{x}}_{BW}$のその他の性質に関する考察，及び具体的な応用場面においての詳細な評価は今後の課題としたい．

謝辞 本研究に関して貴重な御意見を頂きました早稲田大学松嶋研究室，平澤研究室の各位に深く感謝いたします．本研究の一部は，文部省科学研究費基盤 (C)(No.12650400)，早稲田大学特定課題研究助成費 (99A-551) の援助による．

文 献

[1] S.M. Kay, Fundamentals of Statistical Signal Processing, Prentice Hall, New Jersey, 1993.

[2] R.R. Coifman and M.V. Wickerhauser, "Entropy-based algorithms for best basis selection," IEEE Trans. Inf. Theory, vol.38, no.5, pp.713–718, March 1992.

[3] J.C. Pesquet, H. Krim, and D. Leporini, "Bayesian approach to best basis selection," ICASSP, pp.2634–2637, Atlanta, GA, May 1996.

[4] D. Leporini, J.C. Pesquet, and H. Krim, "Best basis representations with prior statistical models," Lecture Notes in Statistics, pp.155–172, Springer-Verlag, New York, 1999.

[5] 北原正樹，野村 亮，松嶋敏泰，"ウェーブレット・パケットを用いた雑音除去におけるベイズ法の応用に関する一考察，" 信学技報，DSP2000-134, Dec. 2000.

[6] H.A Chipman, E.D. Kolaczyk, and R.E. McCulloch, "Adaptive Bayesian wavelet shrinkage," J. American Statistical Association, vol.92, no.440, pp.1413–1421, Dec.1997.

[7] B. Vidakovic, "Nonlinear wavelet shrinkage with Bayes rules," J. American Statistical Association, vol.93, no.441, pp.173–179, March 1998.

[8] N. Saito, "Simultaneous Noise Suppression and Signal Compression Using a Library of Orthormal Bases and the Minimun Description Length Criterion," in Wavelets in Geophysics, pp.299–324, Academic, New York, 1994.

[9] J.O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer-Verlag, Berlin, 1985.

[10] 芦野隆一，山本鎮男，ウェーブレット解析，共立出版，東京，1997.

[11] G. Strang, "Wavelets and dilation equations," SIAM Review, vol.31, no.4, pp.614–627, 1989.

[12] T. Matsushima and S. Hirasawa, "A Bayes coding using context trees," IEEE International Symposium on Information Theory, p.386, 1994.

[13] 後藤正幸，ベイズ統計理論に基づく確率モデルの推定と予測の漸近的評価に関する研究，早稲田大学博士論文，2000.

[14] E.L. Lehmann and G. Casella, Theory of Point Estimation, Springer-Verlag, Berlin, 1998.

[15] B.S. Clarke and A.R. Barron, "Infomation theoretic asymptotics of Bayes methods," IEEE Trans. Inf. Theory, vol.36, no.3, pp.453–471, May 1990.

[16] M. Goto, T. Matsushima, and S. Hirasawa, "An analysis of the difference of code length between two-step codes based on MDL principle and Bayes codes," IEEE Trans. Inf. Theory, vol.47, no.3, pp.927–944, March 2001.

[17] M. Goto, T. Matsushima, and S. Hirasawa, "A generalization of B.S. Clarke and A.R. Barron's asymptotics of Bayes codes for FSMX sources," IEEE Trans. Inf. Theory, vol.47, no.3, pp.2123–2132, March 2001.

付　　録

1. 補題 3.1 の証明

ベイズリスクを最小化する推定量を求めることは，事後期待損を最小化する推定量を求めることと等価である [9]．式 (9) のベイズリスク $BR(g,P,\hat{\boldsymbol{x}})$ に対応した事後期待損 $post(g,P,\hat{\boldsymbol{x}}|\boldsymbol{y})$ は次式で与えられる．

$$\begin{aligned} &post(g,P,\hat{\boldsymbol{x}}|\boldsymbol{y}) \\ &=\sum_{B_i\in L_0}\int_{\boldsymbol{R}^N}\|\boldsymbol{X}_i-\hat{\boldsymbol{X}}_i\|^2 g(\boldsymbol{X}_i|\boldsymbol{Y}_i,B_i)d\boldsymbol{X}_i \\ &\quad\times P(B_i|\boldsymbol{Y}_i) \end{aligned} \tag{A·1}$$

ここで，$\boldsymbol{W}_i$ の n 番目の列ベクトルを $\boldsymbol{w}_{i,n}$ とし，$\hat{\boldsymbol{x}}$ の n 番目の要素を $\hat{x}_n$ とする．すると，$\boldsymbol{W}_i$ は直交行列なので，$post(g,P,\hat{\boldsymbol{x}}|\boldsymbol{y})$ の $\hat{x}_n$ に関する偏微分 $\frac{\partial post(g,P,\hat{\boldsymbol{x}}|\boldsymbol{y})}{\partial \hat{x}_n}$ は次式で与えられる．

$$\begin{aligned} &\frac{\partial post(g,P,\hat{\boldsymbol{x}}|\boldsymbol{y})}{\partial \hat{x}_n} \\ &=2\hat{x}_n-2\sum_{B_i\in L_0}\boldsymbol{w}_{i,n}^T\int_{\boldsymbol{R}^N} g(\boldsymbol{X}_i|\boldsymbol{Y}_i,B_i)\boldsymbol{X}_i d\boldsymbol{X}_i \\ &\quad\times P(B_i|\boldsymbol{Y}_i) \end{aligned} \tag{A·2}$$

$\frac{\partial post(g,P,\hat{\boldsymbol{x}}|\boldsymbol{y})}{\partial \hat{x}_n}=0$，$n\in\{0,1,\cdots,N-1\}$ とおくことにより，$\hat{\boldsymbol{x}}_{BW}$ が導かれる．□

2. 定理 4.1 の証明

帰納法を用いてアルゴリズムの正当性を証明するため，いくつかの新たな定義をしておく．

ここで，(j,k) を根ノードとした WP 木の部分木の葉ノードに対応した基底の和集合によって与えられる基底 $B_i^{j,k}$，$i\in\{0,1,\cdots,M\}$ を考える(注11)．ただし，$B_i^{j,k}$ に対応した部分木の葉ノードのインデックスの集合を $S_{leaf}^{i(j,k)}$，内節ノードのインデックスの集合を $S_{inter}^{i(j,k)}$ とし，

$$V_{j,k}=\oplus_{(l,m)\in S_{leaf}^{i(j,k)}}V_{l,m}, \tag{A·3}$$

が成立するとする．また，$B_0^{j,k}=b_{j,k}$ とし，$B_i^{j,k}$，$i\in\{1,2,\cdots M\}$ の集合を $L_{j,k}$ とする（$L_{0,0}=L_0$ であることに注意する）．

ここで，$\boldsymbol{X}_i^{j,k}$，$\boldsymbol{Y}_i^{j,k}$ を $B_i^{j,k}$ の基底成分に対応した $\boldsymbol{x}$，$\boldsymbol{y}$ の変換係数ベクトルとして表し，

$$g(\boldsymbol{X}_i^{j,k}|B_i^{j,k})=\prod_{(l,m)\in S_{leaf}^{i(j,k)}}g(\boldsymbol{X}_{l,m}|b_{l,m}), \tag{A·4}$$

$$f(\boldsymbol{Y}_i^{j,k}|B_i^{j,k})=\prod_{(l,m)\in S_{leaf}^{i(j,k)}}f(\boldsymbol{Y}_{l,m}|b_{l,m}), \tag{A·5}$$

$$P(B_i^{j,k})=\prod_{(l,m)\in S_{leaf}^{i(j,k)}}u_{l,m}\prod_{(p,q)\in S_{inter}^{i(j,k)}}(1-u_{p,q}), \tag{A·6}$$

とおく．

なお，以下に示す二つの補題が証明されれば，定理 4.1 が証明されることは自明である．

[補題 2.1]　$0\leqq j\leqq J$，$0\leqq k\leqq 2^j$ に関して次式が成立する．

$$P_{j,k}=\sum_{B_i^{j,k}\in L_{j,k}}f(\boldsymbol{Y}_i^{j,k}|B_i^{j,k})P(B_i^{j,k}) \tag{A·7}$$

（証明）　帰納法により式 (A·7) を証明する．$j=J$ の場合は明らかに式 (A·7) が成り立つ．次に $0\leqq j<J$ のときに式 (A·7) が成立しているとする．仮説より，式 (22) は以下のように展開できることより補題が証明される．

$$\begin{aligned} P_{j-1,k}&=u_{j-1,k}f(\boldsymbol{Y}_{j,k}|b_{j,k})+(1-u_{j-1,k}) \\ &\quad\times\left(\sum_{B_i^{j,2k}\in L_{j,2k}}f(\boldsymbol{Y}_i^{j,2k}|B_i^{j,2k})P(B_i^{j,2k})\right. \\ &\quad\left.\times\sum_{B_i^{j,2k+1}\in L_{j,2k+1}}f(\boldsymbol{Y}_i^{j,2k+1}|B_i^{j,2k+1})P(B_i^{j,2k+1})\right) \end{aligned}$$

(注 11)：このようにして得られる基底の数を M と表した．

$$= u_{j-1,k} f(\boldsymbol{Y}_{j,k}|b_{j,k}) + \sum_{B_i^{j,k} \in L^-_{j,k-1}} f(\boldsymbol{Y}_i^{j-1,k}|B_i^{j-1,k}) P(B_i^{j-1,k})$$

$$= \sum_{B_i^{j,k} \in L_{j,k-1}} f(\boldsymbol{Y}_i^{j-1,k}|B_i^{j-1,k}) P(B_i^{j-1,k}) \quad \text{(A·8)}$$

ただし，$L^-_{j,k-1}$は$L_{j,k-1}$から$B_0^{j,k}$を除いた集合である． □

[補題 2.2] $0 \leqq j \leqq J$，$0 \leqq k \leqq 2^j$に関して次式が成立する．

$$\bar{\boldsymbol{X}}_{j,k} = \sum_{B_i^{j,k} \in L_{j,k}} f(\boldsymbol{Y}_i^{j,k}|B_i^{j,k}) P(B_i^{j,k}) \bar{\boldsymbol{X}}_i^{j,k} \quad \text{(A·9)}$$

ただし，$\bar{\boldsymbol{X}}_i^{j,k}$は$b_{j,k}$の変換領域における$\boldsymbol{X}_i^{j,k}$の事後平均ベクトルである．

(証明) 帰納法を用いて式(A·9)を証明する．$j = J$の場合は明らかに式(A·9)は成り立つ．次に$0 \leqq j < J$のときに式(A·9)が成立しているとする．式(23)は仮説，及び式(A·7)が証明されたことから以下のように展開できる．

$$\begin{aligned}
&\bar{\boldsymbol{X}}_{j-1,k} \\
&= u_{j-1,k} f(\boldsymbol{Y}_{j-1,k}|b_{j-1,k}) \bar{\boldsymbol{X}}_{j-1,k} \\
&\quad + (1-u_{j-1,k})(P_{j,2k+1} \bar{\boldsymbol{X}}_{j,2k}^{\uparrow 0} + P_{j,2k} \bar{\boldsymbol{X}}_{j,2k+1}^{\uparrow 1}) \\
&= u_{j-1,k} f(\boldsymbol{Y}_{j-1,k}|b_{j,k}) \bar{\boldsymbol{X}}_{j-1,k} + (1-u_{j-1,k}) \\
&\quad \times \Bigg(P_{j,2k+1} \sum_{B_i^{j,2k} \in L_{j,2k}} f(\boldsymbol{Y}_i^{j,2k}|B_i^{j,2k}) \\
&\quad \times P(B_i^{j,2k}) \bar{\boldsymbol{X}}_i^{j,2k\uparrow 0} \\
&\quad + P_{j,2k} \sum_{B_i^{j,2k+1} \in L_{j,2k+1}} f(\boldsymbol{Y}_i^{j,2k+1}|B_i^{j,2k+1}) \\
&\quad \times P(B_i^{j,2k+1}) \bar{\boldsymbol{X}}_i^{j,2k+1\uparrow 1} \Bigg) \\
&= u_{j-1,k} f(\boldsymbol{Y}_{j-1,k}|b_{j-1,k}) \bar{\boldsymbol{X}}_{j-1,k} + (1-u_{j-1,k}) \\
&\quad \times \Bigg\{ \sum_{B_i^{j,2k} \in L_{j,2k}} \sum_{B_i^{j,2k+1} \in L_{j,2k+1}} f(\boldsymbol{Y}_i^{j,2k}|B_i^{j,2k}) \\
&\quad \times P(B_i^{j,2k}) f(\boldsymbol{Y}_i^{j,2k+1}|B_i^{j,2k+1}) P(B_i^{j,2k+1}) \\
&\quad \times (\bar{\boldsymbol{X}}_i^{j,2k\uparrow 0} + \bar{\boldsymbol{X}}_i^{j,2k+1\uparrow 1}) \Bigg\}
\end{aligned} \quad \text{(A·10)}$$

更に，式(A·7)の証明の場合と同様の原理により以下のように展開され，補題が証明される．

$$\begin{aligned}
&\bar{\boldsymbol{X}}_{j-1,k} \\
&= u_{j-1,k} f(\boldsymbol{Y}_{j-1,k}|b_{j-1,k}) \bar{\boldsymbol{X}}_{j-1,k} + (1-u_{j-1,k}) \\
&\quad \times \sum_{B_i^{j-1,k} \in L^-_{j-1,k}} f(\boldsymbol{Y}_i^{j-1,k}|B_i^{j-1,k}) P(B_i^{j-1,k}) \bar{\boldsymbol{X}}_i^{j-1,k} \\
&= \sum_{B_i^{j-1,k} \in L_{j-1,k}} f(\boldsymbol{Y}_i^{j-1,k}|B_i^{j-1,k}) P(B_i^{j-1,k}) \bar{\boldsymbol{X}}_i^{j-1,k}
\end{aligned} \quad \text{(A·11)}$$

□

(平成13年4月2日受付，10月16日再受付，
14年1月28日最終原稿受付)

北原　正樹

平11早大・理工・経営システム卒．平13同大大学院修士課程了．同年，日本電信電話（株）に入社．在学中，信号処理における統計的推定の適用に関する研究に従事．

野村　亮　（学生員）

平8早大・理工・工業経営卒．平10同大大学院修士課程了．現在，同大学院博士後期課程在学中．情報源符号化に関する研究に従事．IEEE，情報理論とその応用学会各会員．

松嶋　敏泰　（正員）

昭53早大・理工・工業経営卒．昭55同大大学院修士課程了．同年，日本電気（株）入社．昭61早大・理工学研究科・博士後期課程入学．平1横浜商科大学講師．平3同大助教授．平4早大・理工学部・工業経営学科（現在経営システム工学科）助教授，平9同大教授，現在に至る．知識情報処理及び情報理論とその応用に関する研究に従事．工博．IEEE，情報理論とその応用学会，人工知能学会，情報処理学会，OR学会，日本経営工学会等各会員．

論　文

メモリ量を低減した近似ベイズ符号化アルゴリズム

野村　亮[†]　　松嶋　敏泰[†]　　平澤　茂一[†]

An approximation algorithm of Bayes coding to reduce memory capacity

Ryo NOMURA[†], Toshiyasu MATSUSHIMA[†], and Shigeichi HIRASAWA[†]

あらまし　情報源の確率モデルは既知であるが，そのパラメータは未知である場合の符号化法において，ベイズ符号はベイズ基準のもとで冗長度を最小にする符号である．また，ベイズ符号を構成するアルゴリズムとしてFSMX 情報源に対する文脈木を用いたベイズ符号化法が提案されている．このアルゴリズムは文脈木を逐次的に生成することにより，最大深さが任意の FSMX 情報源に対してベイズ符号を構成している．しかし，実用化を考えた場合，系列長とともに文脈木を生成することはメモリの点から困難である．本研究ではメモリ容量を低減した近似ベイズ符号化アルゴリズムを提案しその性能を評価する．

キーワード　ベイズ符号，ユニバーサル符号，文脈木，事後分布

1. はじめに

情報源の確率構造について完全な情報が得られていない場合の符号化法，すなわちユニバーサル情報源符号化法に関しては従来より多くの研究がなされている[1][2]．情報源の分布のクラスのみを仮定し，そのパラメータに関しては未知の場合を扱うユニバーサル符号の中で，ベイズ符号は冗長度をベイズ基準のもとで最小にする符号である[3]．そして，FSMX 情報源に対して文脈木を用いたベイズ符号の効率的なアルゴリズムが提案されている[4][5]．これらのアルゴリズムはあらかじめ文脈木の深さを設定しており，FSMX 情報源の一部のクラスに対するベイズ符号を構成するものであった．一方近年，文脈木を逐次的に生成することによりすべての FSMX 情報源のクラスを対象としたベイズ符号化アルゴリズムが提案されている[6][7]．しかし，これらのアルゴリズムにおいては文脈木のノード数が系列長とともに増大するため，実用化の際には莫大なメモリ量を必要としてしまう．情報源の確率構造を全く仮定しないユニバーサル符号である Ziv-Lempel（LZ）符号[8] は漸近的に最良な符号であるが，じつは同様の問題を抱えており，実用化に際してはある深さまでしか文脈木を成長させない，などのアルゴリズムでこの問題を回避している．

本研究では，文脈木のノード数をメモリ量と考え，まずノード数を低減したアルゴリズムを提案する．つぎに，提案アルゴリズムを用いたときの 1 シンボルあたりの符号長がベイズ符号化法の符号長と漸近的に一致することを示す．さらに，提案アルゴリズムの有限時点での性能をいくつかの数値実験により評価する．

2. ベイズ符号化法

2.1 FSMX 情報源

ベイズ符号は情報源の分布のクラスのみが既知であり，そのパラメータが未知の場合を対象とした符号である．松嶋らにより提案されたベイズ符号を構成する効率的なアルゴリズム（以下，ベイズ符号化法と呼ぶ．）は有限アルファベット上の FSMX 情報源を対象としている．次に FSMX 情報源について定式化を行う．

FSMX 情報源とは過去の有限系列から現在のシンボルの発生確率の決まる情報源でマルコフ過程の一種である．FSMX 情報源は階層型モデルであるため，モデルとそのモデルのもとでのパラメータの二つにより定まる．$x^n : x_1 x_2 \cdots x_n$ を長さ n の情報源系列とすると，FSMX 情報源における t 時点の状態は情報源系列 x^{t-1} により決まる．FSMX 情報源モデル m における状態の集合を $S(m)$ であるとし，この情報源系列 x^{t-1} から状態 $s \in S(m)$ への写像を $s(x^{t-1})$ とする．ここで，情報源アルファベットを $a \in A = \{a | 0 \leq a \leq l-1\}$

[†] 早稲田大学理工学部経営システム工学科
School of Science and Engineering, Waseda University, 3-4-1 Ohkubo Shinjyuku-ku, Tokyo, 169-8555, JAPAN

図 1 binary-FSMX 情報源モデル（木表現）
Fig. 1 An example of binary-FSMX source model

とすると各状態 s でのシンボルの出現確率は $(l-1)$ 次元パラメータベクトル $\theta^s = \{\theta^s_1, \theta^s_2, \cdots, \theta^s_{l-1}\}$ によって決まる．従って，シンボル x_t の x^{t-1} のもとでの条件付き確率は $P(x_t|\theta^{s(x^{t-1})}, s(x^{t-1}))$ となる．

結局，情報源系列 x^t の発生確率は以下で表される．

$$P(x^t) = P(x_1|\theta^{s(\lambda)}, s(\lambda))P(x_2|\theta^{s(x^1)}, s(x^1)) \cdots P(x_t|\theta^{s(x^{t-1})}, s(x^{t-1})), \quad (1)$$

ここで，$s(\lambda)$ は初期状態を表す．

$l-1$ 次元パラメータベクトル θ^s は各状態 s に対応しているので一つの FSMX 情報源のパラメータを $|S(m)|(l-1)$ 次元パラメータベクトル θ^m で表すことにする．すると，FSMX 情報源はモデル m とそのもとでのパラメータを表す θ^m により定義される．

FSMX 情報源モデル m はまた，完全木で表現することができる．木におけるそれぞれの枝はシンボル $a \in A$ に対応している．また，木における葉ノードから根ノードへの一つのパスをコンテクストもしくはポストフィクスと呼ぶ．木表現におけるそれぞれの葉ノードは状態 s と一対一対応しているので，葉ノードは s と書くことができ，FSMX 情報源モデル m における状態の集合 $S(m)$ は木表現における葉ノードの集合といえる．

図 1 に 2 元 FSMX 情報源モデル m_1 の木表現例を示す．文脈 $x^{t-1} = \cdots 10$ で決定される状態を s_{10} と書くことにすると，系列 $x^5 = 10010$ において $t = 2$ 時点の状態は $s(1) = s_1$，$t = 3$ 時点の状態は $s(10) = s_{10}$，$t = 4$ 時点の状態は $s(100) = s_{00}$ となる．また，$S(m_1) = \{s_1, s_{10}, s_{00}\}$ である．

なお，本稿では 2 元 FSMX 情報源 $(l = 2)$ を考え，パラメータのとりうる範囲は 0 から 1 の開区間，すなわち $\theta^s \in (0, 1)$，とする．

2.2 FSMX 情報源に対するベイズ符号化法

情報源系列の確率を仮定すれば算術符号を用いることにより符号化が可能であるため，ユニバーサル情報源符号化の問題はシンボルの出現確率を決定する問題に帰着する．ベイズ符号はベイズ基準のもとで冗長度を最小にシンボルの出現確率（以下，符号化確率と呼ぶ）を決定する．FSMX 情報源モデル m とそのもとでのパラメータ θ^m が共に未知でありかつ，FSMX 情報源モデル m の事前確率 $P(m)$，モデル m のもとでのパラメータ θ^m の事前確率 $P(\theta^m|m)$ が既知であるとき，ベイズ符号の符号化確率は以下の式で求められる．

[補題 2.1] [3] FSMX 情報源 (m, θ^m) に対するベイズ符号の符号化確率は

$$AP(x_t|x^{t-1}) = \sum_{m \in M} \int_{\theta^m} P(x_t|x^{t-1}, \theta^m, m) P(\theta^m|m, x^{t-1}) P(m|x^{t-1}) d\theta^m, \quad (2)$$

である．ここで，$P(\theta^m|m, x^{t-1})$ は FSMX 情報源 m と x^{t-1} のもとでの θ^m の事後確率，また $P(m|x^{t-1})$ は x^{t-1} のもとでの m の事後確率を表す． □

ベイズ符号化法は上記の符号化確率を効率的に求めるアルゴリズムである．以下，そのアルゴリズムについて説明する．ベイズ符号化法は各時点毎に 1) 文脈木の生成を行い，その文脈木を用いて 2) 符号化確率の計算を行う．最初に，文脈木の生成法について説明する．

文脈木の生成

部分系列 x^j_i を $x^j_i = x_i x_{i+1} \cdots x_j$ とする．t 時点のシンボル x_t に対してそのポストフィクス(注1)に対応するノード $s(\lambda), s(x^t_{t-1}), s(x^t_{t-2}), \cdots, s(x^t_1)$ のうち，既に文脈木に含まれているノードを除いたノードを文脈木に加える． □

t 時点のポストフィクスに対応するノードの集合を S_t と書くことにする．つまり，$S_t = \{s(\lambda), s(x^t_{t-1}), s(x^t_{t-2}), \cdots, s(x^t_1)\}$ となる．

このようにして生成された文脈木は現在得られている系列から考えられる全ての FSMX 情報源の状態の集合となっている．

さらに，松嶋らは以下の式を提案した [5]．

$$P(s) = \sum_{\{m|s \in S(m)\}} P(m). \quad (3)$$

(注1)：$x^5 = 01001$ のとき，$x_5(= 1)$ のポストフィクスは，部分系列の集合 $\{\lambda, 0, 00, 100, 0100\}$ である．ここで λ は空系列を表す．

[注意 2.1] ここで，上記の確率は次の式を満たしている．

$$\sum_{s\in S_t} P(s) = 1. \tag{4}$$

この式は文脈木における一つのノードの事前確率にそのノードを状態として含むFSMX情報源モデルの事前確率を対応させたものである．この式を用いることにより，一つのポストフィクスに属するノードを考えるのみでベイズ符号の符号化確率を計算することができる．(3) 式を用い，さらにパラメータの事前分布にディレクレ分布を仮定すると，符号化確率の計算式は次の補題で与えられる．

符号化確率の計算

[補題 2.2] [7]

$$AP(x_t|x^{t-1}) = \sum_{s\in S_t} P^s(x_t|x^{t-1}, s)P(s|x^{t-1}), \tag{5}$$

ここで，

$$P^s(x_t|x^{t-1}, s) = \frac{n(x_t|x^{t-1}, s) + \beta(x_t|s)}{\sum_{i=0}^{l-1} n(i|x^{t-1}, s) + \beta(s)}, \tag{6}$$

である．また，$\beta(x_t|s)$，$\beta(s)$ はそれぞれ既知のディレクレ分布のパラメータである．また $n(i|x^{t-1}, s)$ は x^{t-1} における状態 s のもとでのシンボル i の発生回数で x^{t-1} により計算される． □

(5) 式における和はt時点のポストフィクスに対してとられる．故に，文脈木と (5) 式を用いれば，符号化確率を計算する際に，t時点のポストフィクスのみを考えればよいことになる．

ここで，ノードの事後確率は以下のように更新される．

$$P(s|x^t) = \frac{q(s|x^t)}{\sum_{s\in S_t} q(s|x^t)} \tag{7}$$

ただし，

$$q(s|x^t) = \begin{cases} \frac{P^s(x_t|x^{t-1},s)P(s|x^{t-1})}{AP(x_t|x^{t-1})}, & \text{if } s \in S_t \\ P(s|x^{t-1}), & \text{if } s \notin S_t \end{cases} \tag{8}$$

である．また，$P(s|x^0) = P(s)$ は事前分布であり，既知の $P(m)$ より計算される．

上記のもとで，次の式が成り立つことに注意されたい．

$$P(s|x^t) = \sum_{\{m|s\in S(m)\}} P(m|x^t). \tag{9}$$

松嶋らにより上記の更新過程の簡略法が提案されている [5]．

2.3 ベイズ符号化法の問題点

前節で述べたとおり，松嶋らにより提案されたベイズ符号化法を用いることにより，符号化確率の計算は非常に効率的に行うことができる．しかし，文脈木の生成において系列長が増加すればするほど，文脈木の深さも増加してしまう．

メモリ容量の面から増加する全てのノードを保持しておくことは非常に困難である．そのため，ベイズ符号はベイズ基準のもとで最適な符号であり，その構成アルゴリズムも提案されているが，このままでは実用化は困難と考えられる．

3. 提案アルゴリズム

3.1 FSMX情報源に対するベイズ符号の性質

提案アルゴリズムの前に，本節でFSMX情報源の性質とFSMX情報源に対するベイズ符号化法の性質を調査する．

FSMX情報源 (m^*, θ^{*m^*}) によりデータが発生しているとする．本研究においてこれらをそれぞれ真のモデル m^*，真のモデルにおける真のパラメータ θ^{*m^*} と呼ぶことにする．さらに，m^* の要素を s^* とし，真のノードと呼ぶことにする．真のモデルにおける真のパラメータ θ^{*m^*} を真のノードごとに対応させた θ^{*s^*} を真のノードにおける真のパラメータと呼ぶことにする．さらに，ノード s' が文脈木においてノード s の子孫に当たる場合，その関係を $s' \gg s$ と書くことにする．

ここで，真のノード s^* の定常分布を $q(s^*)$ と書くことにし，文脈木における各ノード s に対して次の値を定義する．

$$\theta^{*s} = \begin{cases} \theta^{*s^*} & if\ s \gg s^* \\ \sum_{s^*\in\bar{S}^*} \frac{q(s^*)\theta^{*s^*}}{\sum_{s^*\in\bar{S}^*} q(s^*)} & if\ s \ll s^*. \end{cases}$$

ここで，$\bar{S}^*$ は真のノードの集合の中でノード s の子孫にあたるノードの集合を指す．すなわち，$\bar{S}^* = \{s^* : s^* \gg s\}$ である．上記により定義されるパラメータを各ノードにおける真のパラメータと呼ぶことにする．つまり，真のノードより子孫側にあれ

ばそのノードの真のパラメータは真のノードのそれと同じである．また，真のノードより親側にあるノードに対しては，その子孫側に真のノードが複数存在するので，それらを定常分布で荷重平均してやったものを真のパラメータと呼ぶ．これらを各ノードの真のパラメータと呼ぶのは以下が成り立つことによる．
$\forall i \in A$ に対して

$$\lim_{n\to\infty} \frac{n(i|x^n, s)}{\sum_{i=0}^{l-1} n(i|x^n, s)} = \theta_i^{*s} \quad a.s., \tag{10}$$

となる．上式は各ノードにおいてシンボル i の出現頻度が θ_i^{*s} に概収束することを意味する．これより文脈木における任意の FSMX 情報源モデル m に対して，真のパラメータ，$\theta^{*m} = \{\theta^{*s_{i1}}, \cdots, \theta^{*s_{i|S(m)|}}\}$，ここで $s_{i1}, \cdots, s_{i|S(m)|} \in S(m)$，が定義できる．

さらに仮定の前に次を準備する．深さ J の葉ノード s^J を持つ FSMX 情報源モデル m_{s^J} に対して，$S(m_{s^J}) = \{si_1, \cdots, s_{i(|S(m)|-1)}, s^J\}$ であるとし，$J \neq 0$ である場合を考える．ノード s^J に対して，木表現におけるその親ノードを s^{J-1} とすると，2 元 FSMX 情報源モデルを考えているため s^{J-1} を祖先として持つ葉ノードが $S(m_{s^J})$ 内に s^J 以外に一つ以上存在する．このノードの集合を S^J とすると $S(m_{s^{J-1}}) = (S(m_{s^J}) \cup \{s^{J-1}\}) \setminus (\{s^J\} \cup S^J)$ なる FSMX 情報源モデル $m_{s^{J-1}}$ が定義できる．また同様にノード s^{J-1} に対しても $s^{J-1} \neq s_\lambda$ であれば，FSMX 情報源モデル $m_{s^{J-2}}$ が定義できる．このように一つの FSMX 情報源モデルに対して，ある深さ J の葉ノード s^J に注目し上記の操作を繰り返すと $J+1$ 個の FSMX 情報源モデルが定義できる．すると $s^J, \cdots, s^0$ はポストフィクスをなす．すなわち，$s^0 = s_\lambda$ であり，$s^0 \ll s^2 \ll \cdots \ll s^J$ が成立している．次に仮定を述べる．

[仮定 3.1] 葉ノード s^J をもつある FSMX 情報源に対して，$0 \leqq j < i \leqq J$ とすると

$$D\left(P^*_{m^*}; P^*_{m_{s^j}}\right) - D\left(P^*_{m^*}; P^*_{m_{s^i}}\right) > 0, \tag{11}$$

が成立する．ここで，P^*_m は確率分布 $P(X|\theta^{*m}, m)$ を表す．また，$D(P(X); Q(X))$ は確率分布 $P(X)$ と $Q(X)$ の間の KL 情報量である． □

上記の仮定は，ある FSMX 情報源と KL 情報量的に最も近い任意の低次の FSMX 情報源に対して，同様の確率分布を表現可能なより低次の FSMX 情報源が存在しないことを意味している．

[例 3.1] 深さ 2 の 2 元マルコフ情報源を考える．各状態でのシンボルの出現確率が次のような場合，この情報源は仮定 3.1 を満たす．$P(0|00) = 0.2, P(0|10) = 0.3, P(0|01) = 0.4, P(0|11) = 0.5$.

また，最大深さ 2 の FSMX 情報源の各状態でのシンボルの出現確率が次のような場合を考える．$P(0|0) = 0.5, P(0|01) = 0.4, P(0|11) = 0.6$. ここで，$s^2 = s_{11}$ であるとすると，$s^1 = s_1, s^0 = s_\lambda$ である．ここで，$S(m_{s^1}) = \{s_0, s_1\}$ であり，$S(m_{s^0}) = \{s_\lambda\}$ である．また，$P(0|\theta^{*s_0}, s_0) = P(0|\theta^{*s_1}, s_1) = 0.5$, $P(0|\theta^{*s_\lambda}, s_\lambda) = 0.5$ である．故に，

$$D\left(P^*_{m^*}; P^*_{m_{s^1}}\right) - D\left(P^*_{m^*}; P^*_{m_{s^0}}\right) = 0, \tag{12}$$

となるのでこの情報源は仮定 3.1 を満たさない． □

FSMX 情報源に対しては以下の式が成立することが示されている [13].

[性質 3.1] 重複対数の法則が成り立つ．任意の i に対して，

$$\limsup_{n\to\infty} \frac{n(i|x^n, s) - n\theta_i^{*s}}{(2n\sigma \log\log n)^{\frac{1}{2}}} = 1 \quad a.s. \tag{13}$$

σ は FSMX 情報源におけるシンボルの分散とする

上記の性質より次が成り立つ．

[系 3.1] $\hat{\theta}^m(x^n)$ を θ^m の最ゆう推定量とすると，

$$||\hat{\theta}^m(x^n) - \theta^{*m}|| \leq O((\frac{\log\log n}{n})^{\frac{1}{2}}) \quad a.s., \tag{14}$$

が成立する． □

[性質 3.2] パラメータの最ゆう推定量が以下を満たす．

$$P(\hat{\theta^m}(x^n)|m, x^n) = (\frac{n}{2\pi})^{\frac{k_m}{2}} \sqrt{\det I(\hat{\theta}^m(x^n)|m)} + o(n^{\frac{k_m}{2}}) \quad a.s., \tag{15}$$

ここで，$I(\theta^m|m)$ は，パラメータ θ^m の Fisher 情報量行列を表す．すなわち，

$$I(\theta^m|m) = -\lim_{n\to\infty} \frac{1}{n} E \frac{\partial^2 \log P(x^n|m, \theta^m)}{\partial\theta^m \partial\theta^{mT}}, \tag{16}$$

である． □

また，大数の強法則より以下も成り立つ．

[性質 3.3] $\forall\theta^m, m$ に対して

$$\frac{1}{n} \log \frac{P(x^n|m^*, \theta^{*m^*})}{P(x^n|m, \theta^m)} - D(p^{\theta^*}_{m^*} || p^{\theta^m}_m)$$

$$= o(1) \quad a.s., \tag{17}$$

となる．ここで，$D(p_{m^*}^{\theta^*}||p_m^{\theta^m})$ は分布 $p(X|\theta^{m^*}, m^*)$ と分布 $p(X|\theta^m, m)$ の間の KL 情報量を示す． □

上記のもとで後藤らは次の補題を示した [13].

[補題 3.1] [13] 任意の FSMX 情報源 $m \neq m^*$ に対して，適当な事前分布のもとで，

$$\lim_{n\to\infty} \left| \frac{P(x^n|m)}{P(x^n|m^*)} \right| = 0 \quad a.s., \tag{18}$$

が成立する．ただし，m の集合 $\mathcal{M}$ は有限集合．

ここで，上記補題の証明に集合 $\mathcal{M}$ は有限である条件を用いていないことより，つぎの系が成立することに注意されたい．

[系 3.2] 任意の FSMX 情報源 $m \neq m^*$ に対して，適当な事前分布のもとで，

$$\lim_{n\to\infty} \left| \frac{P(x^n|m)}{P(x^n|m^*)} \right| = 0 \quad a.s., \tag{19}$$

が成立する．ただし，m の集合 $\mathcal{M}$ は加算無限集合．

□

また，次の補題も成立する．

[補題 3.2] $s' \ll s^*$（もしくは，$s^* \ll s'$）を満たす s' について，適当な事前分布のもとで，

$$\lim_{n\to\infty} \left| \frac{P(x^n|m_{s'})}{P(x^n|m_s)} \right| = 0 \quad a.s., \tag{20}$$

を満たす $s : s' \ll s \underline{\ll} s^* (s^* \underline{\ll} s \ll s')$ が存在する．ただし $s \underline{\ll} s'$ は，$s \ll s'$ または $s = s'$，を意味することとする．

(証明) 付録参照． □

また，次を仮定する

[仮定 3.2] 任意の s に対してその子ノードを s^1 としたとき

$$P(s^1) = \frac{1}{r} P(s), \tag{21}$$

ここで $r > 1$ である． □

上記の仮定は Willems らによる CTW 法においても仮定されている [4].

これらの系，補題より，FSMX 情報源におけるベイズ符号の漸近的性質に関する定理を得る．但し，S を n 時点の文脈木における任意のポストフィクスとする．

[定理 3.1] 任意の FSMX 情報源と任意の $s \neq s^*$ となる $s, s^* \in S$ に対して，

$$\lim_{n\to\infty} \left| \frac{P(s|x^n)}{P(s^*|x^n)} \right| = 0 \quad a.s. \tag{22}$$

が成り立つ．

(証明) 式 (9) より，

$$\frac{P(s|x^n)}{P(s^*|x^n)} \tag{23}$$

$$= \frac{\sum_{\{m|s\in S(m)\}} P(x^n|m)P(m)}{\sum_{\{m|s^*\in S(m)\}} P(x^n|m)P(m)}$$

$$\leqq \frac{\sum_{\{m|s\in S(m)\}} P(x^n|m)P(m)}{P(x^n|m^*)P(m^*)}, \tag{24}$$

である．上式分子の集合は可算無限個の要素を持つ．そこで，次のように分けて考える．

$$\bar{M}_s = \{m|D(P_{m^*}^*; P_m^*) > 0\}, \tag{25}$$

$$\bar{M}_s^C = \{m|D(P_{m^*}^*; P_m^*) = 0\}. \tag{26}$$

ここで，補題 3.2 の証明より，任意の $m \in \bar{M}_s^C$ に対して

$$\lim_{n\to\infty} \frac{p(x^n|m)}{p(x^n|m^*)} = O(\frac{1}{n^{\frac{k_{m^*}-k_m}{2}}}) \tag{27}$$

である．また，任意の $m \in \bar{M}_s^C$ において，$k_m > k_m^*$ である．故に c_i を，$\bar{M}_s^C$ において $k_m = k_{m^*} + i$ であるモデルの数とすると

$$\lim_{n\to\infty} \frac{\sum_{m\in\bar{M}_s^C} P(x^n|m)P(m)}{P(x^n|m^*)P(m^*)}$$

$$= \sum_{i=1}^{\infty} O(\frac{c_i}{n^{\frac{i}{2}}}) \quad a.s., \tag{28}$$

である．ここで，$c_i \leqq |S(m^*)|^i$ であることを考えると，結局

$$\sum_{i=1}^{\infty} O(\frac{c_i}{n^{\frac{i}{2}}}) \leqq \sum_{i=1}^{\infty} O(\frac{|S(m^*)|^i}{n^{\frac{i}{2}}}), \tag{29}$$

が成立する．ここで，

$$\sum_{i=1}^{\infty} \frac{|S(m^*)|^i}{n^{\frac{i}{2}}} = \sum_{i=1}^{\infty} (\frac{|S(m^*)|}{n^{\frac{1}{2}}})^i$$

$$= \sum_{i=1}^{\infty} (\frac{|S(m^*)|}{n^{\frac{1}{2}}})^{i-1} - 1, \tag{30}$$

である．ここで $\sum_{i=1}^{\infty} (\frac{|S(m^*)|}{n^{\frac{1}{2}}})^{i-1}$ は初項 1，項比 $\frac{|S(m^*)|}{n^{\frac{1}{2}}}$ の等比級数で，その和は n が十分大きいと

き，1 に収束する．故に

$$\sum_{i=1}^{\infty}(\frac{|S(m^*)|}{n^{\frac{1}{2}}})^{i-1}-1=0, \tag{31}$$

となる．上式は式 (28) が 0 に収束することを示している．

一方，$\bar{M}_s$ は $D(P^*_{m^*};P^*_m)$ の値により，さらにいくつかの集合に分割することができる．

例えば，

$$S(m^*) \tag{32}$$
$$=\{s(000),s(100),s(10),s(01),s(011),s(111)\}$$

であるとし，$s^*=s(000)$，$s=s(00)$，であるとする．この場合，$S(m)=\{s(00),s(10),s(1)\}$，というモデルと同様の確率分布を表現する集合と，$S(m)=\{s(00),s(10),s(01),s(11)\}$，というモデルと同様の確率分布を表現する集合，{ s(00), s(10),s(01),s(011),s(111)}，と同様の確率分布を表現する集合の三つに分けることができる．

この集合を $\bar{M}_{s1},\bar{M}_{s2},\cdots,\bar{M}_{sj}$，とすると，$j$ は m^* と s に依存するが有限である．ここで，補題 3.2 の証明より，任意の $m\in\bar{si}$ に対して

$$\lim_{n\to\infty}\frac{p(x^n|m)}{p(x^n|m^*)}=O(e^{-nD(P^*_{m^*};P^*_m)})\ \ a.s., \tag{33}$$

であるので，

$$\lim_{n\to\infty}\frac{\sum_{m\in\bar{M}_{si}}P(x^n|m)P(m)}{P(x^n|m^*)P(m^*)}$$
$$=\sum_{i=1}^{\infty}O(\frac{|S(m^*)|^i}{e^{nD(P^*_{m^*};P^*_m)}})\ \ a.s., \tag{34}$$

が成立する．上式は 0 に収束することが分かるので，結局

$$\lim_{n\to\infty}\frac{\sum_{\{m|s\in S(m)\}}P(x^n|m)P(m)}{P(x^n|m^*)P(m^*)}$$
$$=\lim_{n\to\infty}\sum_{i=1}^{j}\frac{\sum_{m\in\bar{M}_{si}}P(x^n|m)P(m)}{P(x^n|m^*)P(m^*)}$$
$$+\frac{\sum_{m\in\bar{M}_s^C}P(x^n|m)P(m)}{P(x^n|m^*)P(m^*)}$$
$$=0\ \ a.s., \tag{35}$$

を得る．これは定理が成り立つことを示している．

□

以上の議論より以下の補題が成り立つ．

［補題 3.3］ $\forall s^*\in S(m)$ に対して仮定 3.2 のもとで

$$\lim_{t\to\infty}\left|P^s(x_t|x^{t-1},s^*(x^{t-1}))-AP(x_t|x^{t-1})\right|$$
$$=0\ \ a.s., \tag{36}$$

が成立する．ここで，$s^*(x^{t-1})$ は x^{t-1} のポストフィクスにおける s^* を表す．□

(証明)

$$\sum_{s:s\neq s^*,s\in S_n}\frac{P(s|x^n)}{P(s^*|x^n)}$$
$$=\sum_{s:s\neq s^*,s\ll s^*,s\in S_n}\frac{P(s|x^n)}{P(s^*|x^n)}$$
$$+\sum_{s:s\neq s^*,s\gg s^*,s\in S_n}\frac{P(s|x^n)}{P(s^*|x^n)} \tag{37}$$

が 0 に収束することを示す．ここで，上式の右辺第一項は有限個の和であるので定理 3.1 より，任意の $\nu>0$ に対して十分大きい n で

$$\sum_{s:s\neq s^*,s\ll s^*,s\in S_n}\frac{P(s|x^n)}{P(s^*|x^n)}<\nu \tag{38}$$

である．

一方，s^* の S_n に含まれる子ノードを s^1，その子ノードを s^2 とし以下，$s^*\ll s^1\ll s^2\ll\cdots\in S_n$ とする．補題 3.2 の証明と同様に考えると，仮定 3.2 より

$$\lim_{n\to\infty}\frac{P(s^{i+1}|x^n)}{P(s^i|x^n)}<O(\frac{1}{r})\ \ a.s., \tag{39}$$

である．ゆえに任意の $\mu>0$ に対して

$$\sum_{s:s\neq s^*,s\gg s^*,s\in S_n}\frac{P(s|x^n)}{P(s^*|x^n)}<\mu\ \ a.s., \tag{40}$$

となるように n を選ぶことができる．

定理 3.1 と上記の議論から任意の $\epsilon>0$

$$\sum_{s:s\in S_n,s\neq s^*}\frac{P(s|x^n)}{P(s^*|x^n)}$$
$$=\mu+\nu<\epsilon, \tag{41}$$

となる n が存在する．すなわち

$$\lim_{n\to\infty}P(s^*|x^{t-1})=1\ \ a.s., \tag{42}$$

が成立することがわかる．

上記は補題が成立することを意味している． □

上の定理は真のノードの事後確率が漸近的に1に概収束することを示している．つまり，漸近的には真のノードのみから計算した符号化確率とベイズ符号の符号化確率とは一致することがわかる．

3.2 提案アルゴリズム

前節で見たように真のノードのみを考慮すれば漸近的にベイズ符号と一致する符号化確率を得ることができる．言い換えると，事後確率最大ノードのみを考えればよいということである．しかしながら，事後確率最大ノードを探索するためには全てのノードを保持しておかねばならず，ベイズ符号化法と使用するノード数は変わらない．そこで，本研究では一つのポストフィクスに対して有限個のノードを用い，漸近的に事後確率最大のノードを含むアルゴリズムを提案する．アルゴリズムは符号化確率の計算，事後確率の更新，文脈ノード集合の更新の三つの過程からなる．ベイズ符号化法と大きく異なるのは文脈ノード集合の更新方法であり，その文脈木のもとでは符号化確率の計算と事後確率の更新はベイズ符号化法とほぼ同様である．詳しい文脈ノード集合の更新方法は付録に示し，ここでは概要のみを述べる．なお，提案するノード集合は木ではなくグラフとなっている．それゆえ，本研究ではこのノード集合のことを文脈グラフと呼ぶ．

準備　提案アルゴリズムは一つのポストフィクスに対して $C(>1)$ 個のノードのみを用いる．$M(x^{t-1})$ を $t-1$ 時点において提案アルゴリズムにより作られた文脈グラフとする．$M(x^{t-1})$ は x^{t-1} により定まる確率変数である．そして，$S_p(x^{t-1})$ を提案アルゴリズムにより作られる文脈グラフに含まれるノードと x_t のポストフィクスに対応するノードの共通集合とする．すなわち，$S_p(x^{t-1}) = \{s|M(x^{t-1}) \cap S_t\}$ であり，$|S_p(x^{t-1})| = C$ である．またしきい値 $\alpha, \beta > 0$ を用意する．ここでは，x^{t-1} の条件付符号化確率 $AP_p(x_{t-1}|x^{t-2})$ まで計算済みであり，$M(x^{t-1})$ は得られているものとする．このもとで $AP_p(x_t|x^{t-1})$ の計算法，提案文脈グラフの更新法を述べる．なお，以下では $M(\cdot)$ における任意のポストフィクス $S_p(x^i)$ に対して，$\max_j s(x^i_{i-j})$ を満たすノードを葉ノードと呼ぶことにする．さらに，従来の文脈木と同様にノード間の親子関係を次のように定義する．すなわち，任意のポストフィクス $S_p(x^i)$ において $s(x^i_j), s(x^i_{j+1}) \in S_p(x^i)$ となる場合 $s(x^i_j)$ を $s(x^i_{j+1})$ の子ノードと呼ぶことにする．

1. 符号化確率の計算

x^t と $M(x^{t-1})$ から $S_p(x^{t-1})$ を求め，$S_p(x^{t-1})$ における事後確率を正規化する．

$$P_p(s|x^{t-1}) = \frac{P_p(s|x^{t-1})}{\sum_{s\in S_p(x^{t-1})} P_p(s|x^{t-1})}, \qquad (43)$$

正規化後の $P_p(s|x^{t-1})$ と次式を用いて符号化確率 $AP_p(x_t|x^{t-1})$ を計算する．

$$AP_p(x_t|x^{t-1}) = \sum_{s\in S_p(x^{t-1})} P^s(x_t|x^{t-1}, s)P_p(s|x^{t-1}), \qquad (44)$$

ここで，$P^s(x_t|x^{t-1}, s)$ は (6) 式で定義される．

2. 事後確率の更新

次式を用いて $M(x^{t-1})$ におけるノードの事後確率を更新する．

$$P_p(s|x^t) = \frac{q_p(s|x^t)}{\sum_{s\in S_p(x^{t-1})} q_p(s|x^t)}, \qquad (45)$$

ここで

$$q_p(s|x^t) = \begin{cases} \frac{P^s(x_t|x^{t-1},s)P_p(s|x^{t-1})}{AP_p(x_t|x^{t-1})}, & \text{if } s \in S_p(x^{t-1}) \\ P_p(s|x^{t-1}), & \text{if } s \notin S_p(x^{t-1}) \end{cases} \qquad (46)$$

である．

3. 文脈グラフの更新

ステップ1　$M(x^t) = M(x^{t-1})$ とする．$S_p(x^{t-1})$ における葉ノードを $s(x^{t-1}_j)$ とし，その親ノードを $s(x^{t-1}_{j+1})$ とする．$\frac{P_p(s(x^{t-1}_{j+1})|x^t)}{P_p(s(x^{t-1}_j)|x^t)}$ を計算し，α よりも小さければ，ノード $s(x^{t-1}_{j+C-1})$ を $M(x^t)$ から削除する．さらに $s(x^{t-1}_{j+C-1})$ の子ノード集合 $\underline{M}(C, s(x^{t-1}_{j+C-1}))$ を $M(x^t)$ に追加する．ここで，$\underline{M}(C, s(x^{t-1}_{j+C-1}))$ は削除するノードより C 世代子孫に当たる全てのノードを表す．すなわち，$|\underline{M}(C, s(x^{t-1}_{j+C-1}))| = 2^C$，である．

ステップ2　追加ノードの $n(i|x^t)$ は系列を遡って数え上げる．そのため，任意の $i \in A$ に対して，$P^s(i|x^t, s)$ は従来のベイズ符号化法と同様の値となる．また，

ステップ3　追加ノードの事後確率は次のようになる．

$$P_p(s(x_{j-1}^{t-1})|x^t) = P^s(x_t|s(x_{j-1}^{t-1}))P(s(x_{j-1}^{t-1})), \quad (47)$$

ここで

$$P^s(x_t|s) = \frac{\Gamma(\beta(s))\Gamma(\beta(s|x^t)+n(0|x^t,s))}{\Gamma(\beta(x^t|s))\Gamma(\beta(s))} \cdot \frac{\Gamma(\beta(s)-\beta(s|x^t)+n(1|x^t,s))}{\Gamma(\sum_{i=0}^{1} n(i|x^t,s)+\beta(s))}, \quad (48)$$

である.

上記のようにして $M(x^t)$ に子孫ノードを付け加えるかどうかを決定する. 一つノードを付け加えるごとにもっとも親ノード側にあるノードを削除するので一つのポストフィクスに対して用いるノードは常に C 個である. さらに, 子ノード側にノードを伸ばすのみでは真のノードより深いノードを選択してしまう場合 (以降 overestimate と呼ぶ) が生じる. ステップ 2 において子孫ノードを付加しなかった場合のみ, 以下のようなアルゴリズムで親ノード側に対しても判定してやることにする.

ステップ 4 $S(x_l^{t-1})$ を $S_p(x^{t-1})$ において最も親ノード側にあるノードとする.

ステップ 5 $\frac{P_p(s(x_{i-1}^{t-1})|x^t)}{P_p(s(x_i^{t-1})|x^t)}$ を計算し, β より小さければノード $s(x_{i+1}^{t-1})$ を $M(x^t)$ に付け加え, $\underline{M}(C, s(x_{i+1}^{t-1}))$ に含まれる全てのノードを $M(x^t)$ から削除する.

ステップ 5 追加ノードの $n(i|x^t)$ は系列を遡って数え上げる. また, 追加ノードの事後確率は

$$P_p(s(x_{i+1}^{t-1})|x^t) = \frac{P^s(x^t|s(x_{i+1}^{t-1}))P(s(x_{i+1}^{t-1}))}{AP_p(x^t)}, \quad (49)$$

ここで,

$$AP_p(x^t) = \Pi_{i=1}^{t} AP_p(x_t|x^{t-1}), \quad (50)$$

であるとする.

[注意 3.1] 符号化確率を計算する際に $M(x^t)$ における任意のポストフィクスにおいて事後確率の和が 1 である必要がある. そのため, 提案アルゴリズムでは符号化確率を計算する前に式 (43) を用いて事後確率を正規化する. これは提案文脈グラフ $M(x^t)$ における任意のポストフィクス S_p に置いて $\sum_{s\in S_p} P_p(s|x^t) = 1$, が成立しているとは限らないからである. これには次のような場合が考えられる.

ノードを文脈グラフに追加しない場合を考える. 式 (45) により x^{t-1} 時点で更新された事後確率に対して $\sum_{s\in S_p(x^{t-1})} P_p(s|x^{t-1}) = 1$, が成立するが $\sum_{s\in S_p(x^t)} P_p(s|x^{t-1}) = 1$, が成立するとは限らない. 例えば, $M(x^{t-1}) = \{s(\lambda), s(0), s(1)\}, C = 2$ の場合を考え, $P_p(s(\lambda)|x^{t-1}) + P_p(s(0)|x^{t-1}) = 1, P_p(s(\lambda)|x^{t-1}) + P_p(s(1)|x^{t-1}) = 1$, であるとする. ここで $x^{t-1} = \cdots 00$ であった場合, $S_p(x^{t-1}) = \{s(\lambda), s(0)\}$, であり, 符号化確率は次のようになる.

$$AP_p(x_t|x^{t-1}) = P^s(x_t|s(\lambda))P_p(s(\lambda)|x^{t-1}) + P^s(x_t|s(0))P_p(s(0)|x^{t-1}),$$

各ノードの事後確率は式 (45) より

$$P_p(s(\lambda)|x^t) = \frac{P^s(x_t|s(\lambda))P_p(s(\lambda)|x^{t-1})}{AP_p(x_t|x^{t-1})},$$

$$P_p(s(0)|x^t) = \frac{P^s(x_t|s(0))P_p(s(0)|x^{t-1})}{AP_p(x_t|x^{t-1})},$$

$$P_p(s(1)|x^t) = P_p(s(1)|x^{t-1}),$$

のように更新される. この更新により明らかに $P_p(s(\lambda)|x^t) + P_p(s(1)|x^t) = 1$, は成立していない. 従来のベイズ符号化法においても同様の問題があるが, 従来のベイズ符号化法においては親ノードと子ノードの事後確率の比を保存しておき, 計算量を低減するアルゴリズムが提案されている [5] [6]. 提案アルゴリズムにもこれらの手法が計算量の低減に有効であると考えられる. □

[注意 3.2] 追加ノードの事後確率は式 (47), 式 (49) により決定される. このアルゴリズムを用いるとステップ 1 あるいはステップ 5 の際の事後確率の比較の式において

$$\frac{P_p(s|x^t)}{P_p(s'|x^t)} = \frac{P(s|x^t)}{P(s'|x^t)}, \quad (51)$$

となることに注意されたい. □

[例 3.2] 図 2 に提案文脈グラフの例 ($C = 2$) を示す. 提案文脈グラフは図の白丸ノードと点線部分の枝は保持せず, 黒丸ノードのみを保持している. (注2)現

(注2): 従来のベイズ符号化法における文脈木は白丸, 黒丸と全ての枝を含んだ木になる.

図 2　提案文脈グラフ ($C=2$)
Fig. 2　An example of the proposed context graph

図 3　提案文脈グラフ ($C=2$) 更新後
Fig. 3　An example of the proposed context graph

時点での提案文脈グラフが図 2 のような状態であり，$AP_p(x_t|\cdot 10)$ を計算したいとする．(44) 式に基づき

$$\begin{aligned} &AP_p(x_t|\cdot 10) \\ &= P^s(x_t|x^{t-1}, s(0))P_p(s(0)|x^{t-1}) \\ &\quad +P^s(x_t|x^{t-1}, s(10))P_p(s(10)|x^{t-1}), \end{aligned} \tag{52}$$

を計算し事後確率を更新した後，$\frac{P_p(s(0)|x^t)}{P_p(s(10)|x^t)}$ を計算，α 未満であれば，図3のように文脈グラフを更新する．ここで，$\underline{M}(2, s(0)) = \{s(000), s(100), s(110), s(010)\}$ である．図 2,3 で示すように各ポストフィクスにおいて常に $C=2$ 個のノードが保持されている．すなわち，ここで $t+1$ 時点のポストフィクス S_{t+1} を，$S_{t+1} = \{s(\lambda), s(0), s(10), s(110), \cdots\}$，とすると，$S_p(x^t) = \{s(10), s(110)\}$，である．　□

例で示したように，提案文脈グラフは各ポストフィクスにおいて C 個のノードを保持するアルゴリズムになっている．また，提案文脈グラフは根ノードを持たない場合があるので，一般的に一つの木では表すことはできない．図 3 のように従来のベイズ符号化法における文脈木の部分木を複数含んだノードの集合になっている．

提案アルゴリズムと従来のベイズ符号化法との符号化確率の計算法の相違点は，混合をとるノード集合が異なるという点である．従来法では混合をとる集合が系列長とともに増加していったが，提案法では混合をとる集合は常に定数 C である．

4.　提案アルゴリズムの性能評価

この章では提案アルゴリズムの性能を評価する．

4.1　符号長に関する理論評価

提案アルゴリズムの性能評価に関する主定理は以下のとおり．

[定理 4.1]　仮定 3.1 を満たす任意の FSMX 情報源に対して，次を満たす N_0 が存在する．$\forall n > N_0$ において $s^* \in S_p(x^n)$ がほとんどすべての x^n で成立する．　□

(定理 4.1 の証明) 提案アルゴリズムにおける $P^s(i|x^n, s)$ は系列を遡って数えられていることより，任意の $s', s \in S_p(x^n)$ に対して次の等式が成立する．

$$\begin{aligned} \frac{P_p(s'|x^n)}{P_p(s|x^n)} &= \frac{P(x^n|s')P(s')}{P(x^n|s)P(s)} \\ &= \frac{\sum_{\{m|s'\in S(m)\}} P(x^n|m)P(m)}{\sum_{\{m|s\in S(m)\}} P(x^n|m)P(m)}. \end{aligned} \tag{53}$$

二番目の等式は $P(s|x^n)$ の定義より成立する．

ここで，定理 3.1 の証明と同様に考えると，式 (53) の右辺は 0 に概収束する．故に式 (53) より，$s' \ll s \ll s^*$ であるような任意の $s' \ll s$ に対して一様に次が成立する．すなわち，$s', s \in S_p$ を満たすほとんどすべての x^n に対して

$$\lim_{n\to\infty} \frac{P_p(s'|x^n)}{P_p(s|x^n)} < \alpha \quad a.s., \tag{54}$$

が成立する．同様に式 (53) と補題 4.1 より，$s^* \ll s \ll s'$ であるような任意の $s' \ll s$ に対して一様に次が成立する．すなわち，$s', s \in S_p$ を満たすほとんどすべての x^n に対して

$$\lim_{n\to\infty} \frac{P_p(s'|x^n)}{P_p(s|x^n)} < \beta \quad a.s., \tag{55}$$

が成立する．これより，α, β が 0 より大きければ提案アルゴリズムは確率 1 で s^* を含むことがわかる．故に定理は証明された．　□

さらに，次の補題が成立する．

[補題 4.1] $s, s^* \in S_n$, $\forall s \neq s^*$ を満たす任意の s, s^* に対して仮定 3.1 のもとで，一様に次が成立する．すなわち，$s^*, s \in S_p$ を満たすほとんどすべての x^n に対して

$$\lim_{n\to\infty} \left| \frac{P_p(s|x^n)}{P_p(s^*|x^n)} \right| = 0 \quad a.s., \tag{56}$$

が成立する．

(証明) 定理 4.1 の証明と同様に

$$\frac{P_p(s|x^n)}{P_p(s^*|x^n)} = \frac{P(x^n|s)P(s)}{P(x^n|s^*)P(s^*)} = \frac{\sum_{\{m|s\in S(m)\}} P(x^n|m)P(m)}{\sum_{\{m|s^*\in S(m)\}} P(x^n|m)P(m)}, \tag{57}$$

ここで，$s^* \in S(m^*)$ であることを考えると，式(57)は，

$$\frac{P_p(s|x^n)}{P_p(s^*|x^n)} \leqq \frac{\sum_{\{m|s\in S(m)\}} P(x^n|m)P(m)}{P(x^n|m^*)P(m^*)} \tag{58}$$

が成立する．ここで定理 3.1 と同様に考えると上式右辺は 0 に収束する．故に補題が証明された． □

定理 4.1 と補題 4.1 より提案アルゴリズムを用いたときの符号化確率が漸近的にベイズ符号化法の符号化確率と一致することがわかる．すなわち次の定理が成り立つ．

[定理 4.2] 仮定 3.1，3.2 のもとで

$$\lim_{t\to\infty} -\log \frac{AP_p(x_t|x^{t-1})}{AP(x_t|x^{t-1})} = 0 \quad a.s., \tag{59}$$

が成り立つ． □

(定理 4.2 の証明) 定理 4.1 より，提案アルゴリズムを用いたとき S_p は確率 1 で s^* を含むことがわかる．このことと補題 4.1，$|S_p| = \mathrm{C}$ であることより，

$$\lim_{t\to\infty} P_p(s^*|x^{t-1}) = 1 \quad a.s., \tag{60}$$

であることがわかる．故に，(44) 式より

$$\lim_{t\to\infty} \left| AP_p(x_t|x^t - 1) - P^s(x_t|s^*(x^{t-1}), x^{t-1}) \right| = 0 \quad a.s., \tag{61}$$

を得る．これより定理は証明された． □

上記の定理は提案アルゴリズムのシンボルあたりの符号長が漸近的にベイズ符号化法のシンボルあたりの符号長と一致することを示している．

また漸近的には，しきい値 α, β は $0 < \alpha, \beta < \infty$ の間の任意の値でよいこと，$C > 1$ の個数も任意でよいことが定理の証明からわかる．

4.2 メモリ容量・計算量に関する評価

提案アルゴリズムの部分で述べたように提案文脈グラフは各ポストフィクスにおいて C 個のノードのみを保持している．また，前節で示したように十分大きい n のもとで，各ポストフィクスにおいて提案文脈グラフは真のノードを確率 1 で含む．故に FSMX 情報源 (m^*, θ^{*m^*}) に対して提案アルゴリズムは高々 $C|S(m^*)| + \epsilon$ 個のノードを保持しておけば良い．ここで，ϵ は overestimate してしまった際に余分に必要になるノードの個数を表している．

一方，従来のベイズ符号化法は逐次的に文脈木を生成していくため，必要なノードの個数は系列長 n の指数オーダーとなる．

計算量の面を考えると，従来のベイズ符号化法は混合をとる集合が増加するために計算量も系列長 n とともに増加していく．その量は $O(n)$ である．一方，提案アルゴリズムはノードを付け加えるかどうかの判定の分の計算量が $O(1)$，符号化確率 $AP_p(x_n|x^{n-1})$ を計算するための計算量は C のみに依存して，n には依存しないので，$O(1)$ である．さらにノードを付け加える際の $P^s(x_n|s, x^{n-1})$ を計算する為の計算量は過去の系列 x^{n-1} を一度参照し，数え上げるので $O(n)$ である．これらをすべて加えると提案アルゴリズムに必要な計算量は高々 $O(n)$ であることがわかる．

つまり，提案アルゴリズムは従来法に比べ，オーダーで考えると計算量は同等でメモリ容量が少なくてすむことがわかる．

また，FSMX モデルを選択するという観点から言うと，提案アルゴリズムは Rissanen の提唱した記述長最小基準 [9] に基づく符号 (以下 MDL 符号) とも関係が深い．ここで，MDL 符号もモデルを探索する際に全てのノードを保持しておく必要があり，ベイズ符号と同等のメモリ量が必要になる．そのため，本研究と同様の問題設定に対してはメモリ量が系列長と共に指数的に増加してしまうという問題があることに注意されたい．

4.3 数値実験による評価

4.3.1 実験条件

前節で示したように漸近的には提案アルゴリズムを用いた際の符号長はベイズ符号の符号長と一致する．本節では有限時点での提案アルゴリズムの性能を数値実験により評価する．実験には最大深さ 9 の binary-FSMX 情報源（状態数 28）を用い，50 系列に対する一文字あたりの符号長の平均を取った．実験 1

では提案アルゴリズムの符号長を評価する実験を行った．この結果を図4に示す．Bayesはベイズ符号化法，MDLはMDL符号，$C=3$は$\alpha=\beta=0.3, C=3$の場合を表している．

実験2ではαとβの違いに提案アルゴリズムの性能を見た．使用した情報源は実験1と同様である．この情報源に対して1000時点毎に真のノード28個のうちいくつが提案文脈グラフに含まれるかを調べた．条件は$\alpha=\beta=0.1$，$\alpha=\beta=0.3$，$\alpha=\beta=0.5$である．この結果を図5に示す．

4.3.2 実験結果に関する考察

実験結果1より，提案アルゴリズムを用いた場合の一文字あたりの符号長が漸近的にベイズ符号の一文字あたりの符号長に近づくことが確認できた．また，提案手法はMDL符号とベイズ符号の中間の性能を示している．実際にはCやα,βの値を変更していくつか実験を行ったが，ベイズ符号に近い性能を示すかMDL符号により近い性能を示すかの違いはあったが，全て同様の結果を示した．これは，提案アルゴリズムが使用するモデルの点からはMDL符号とベイズ符号の中間に位置することによると考えられる．また，ベイズ符号，MDL符号を実行するために必要なメモリ量は入力系列長に対し指数的に増加していくことを考えると，この結果は非常に意味がある結果だと考えられる．

実験結果2に対しては，理論上はαもβも正の値であれば漸近的に性能は変わらない．しかしながら，その違いは符号長の収束速度に関係していると考えられる．結果より，系列数が短い時点ではα，βが大きい値のほうがより多く真のノードを含んでいることがわかる．しかし，α，βが大きい場合，系列長が長くななっても個数があまり増えていない．これはα，βが大きい場合，overestimateしてしまう可能性が非常に大きくなるためだと考えられる．α，βの値に関しては符号長によってα,βの値を変化させることも考えると非常に多くのバリエーションが考えられるが，本研究ではこれについては言及しない．

図4 提案アルゴリズムとベイズ符号の平均符号長
Fig. 4 codelength of the proposed algorithm and Bayes code

図5 提案文脈グラフに含まれる真のノード数
Fig. 5 number of true nodes included in the proposed context graph

5. まとめ

本研究ではメモリ容量を低減したベイズ符号化法の近似計算アルゴリズムを提案した．さらに，提案アルゴリズムのシンボルあたりの符号長がベイズ符号化法の符号長と一致することを示した．ベイズ符号化法はベイズ基準のもとで冗長度を最小にする符号であるが，全ての候補ノードを用い符号化確率を求めるため，実用化にはメモリ容量，計算量の点を低減する必要がある．そのため，本研究のようにメモリ容量を低減した近似的なベイズ符号化アルゴリズムは実用化のために意味があると思われる．

謝辞 本研究に関して貴重なご意見を頂きました神奈川工科大学新家稔央氏，武蔵工業大学後藤正幸氏，ならびに早稲田大学松嶋研究室，平澤研究室の各位に深く感謝いたします．本研究の一部は，文部省科学研究費基盤(C)(No.12650400)，早稲田大学特定課題研究助成費(2001A-570)の援助による．

文　献

[1] L.D.Davisson, "Universal Noiseless Coding." IEEE Trans. Inf. Theory, vol.19,no.6,pp.783-795,1973.

[2] J.Ziv and H.Lempel, "A universal algorithm for sequential data compression." IEEE Trans. Inf. Theory, vol.23,no.3,pp.337-343,1977.

[3] T.Matsushima, H.Inazumi and S.Hirasawa, "A Class of Distortionless Codes Designed by Bayes Decision Theory." IEEE Trans. Inf. Theory, vol.37, no.5, pp.1288-1293, 1991.

[4] F.M.J. Willems, Y.M. Shtarkov and T.J. Tjalkens, "The context tree weighting method: Basic properties." IEEE Trans. Inf. Theory,vol.41,no.3,pp.653-663,1995.
[5] T.Matsushima, and S.Hirasawa, "A bayes coding using context tree." In Proc.Int. Symp. on. Inf. Theory ,page 386,1994.
[6] T.Kawabata and F.M.J.Willemes,"A Context Tree Weighting Algorithm with an Incremental Context Set." IEICE Trans. Fundamentals, vol.E83-A, no.10, pp,1898-1903, 2000.
[7] T.Matsushima and S.Hirasawa, "A Bayes coding Algorithm for Markov models." IEICE Technical Report, IT95-1,1995.
[8] J.Ziv and H.Lempel, "Compression of Individual Sequences via Variable-Rate Coding",IEEE Trans. Inf. Theory,vol.24,no.5,pp.530-536,1978.
[9] J.Rissanen, "Modeling by shortest data description", Automatica, vol.46, pp.465-471, 1978.
[10] W.Feller, "確率論とその応用" 紀伊国屋書店, 1960.
[11] 韓 太舜, 小林 欣吾, "情報と符号化の数理" 岩波書店, 1994.
[12] B.S.Clarke and A.R.Barron, "Information Theoretic Asymptotics of Bayes Methods",IEEE Trans. Inf. Theory,vol.36,no.3,pp453-471,1990.
[13] M.Gotoh, T.Matsushima and S.Hirasawa, "A Generalization of B.S.Clarke and A.R.Barron's Asymptotics of Bayes Codes for FSMX Sources" IEICE Trans. Fundamentals, vol.E81-A, no.10,1998.

付　　録

1. 提案文脈木更新アルゴリズム

begin

$M(x^t) = M(x^{t-1})$

$j := \arg\min_j s(x_j^{t-1}) \in M(x^t)$;

$nc = C$;

/* A decision to add child node */

if $\frac{P_p(s(x_{j+1}^{t-1})|x^t)}{P_p(s(x_j^{t-1})|x^t)} < \alpha$ **then**

add $\underline{M}(C, s(x_{j+C-1}^{t-1}))$ *to* $M(x^t)$

delete $s(x_{j+C-1}^{t-1})$ *from* $M(x^t)$;

else

while $nc \geqq 0$ **do**

$j := j - 1$;

$nc := nc - 1$;

end-while

/* A decision to add ancestor node */

if $\frac{P_p(s(x_{j-1}^{t-1})|x^t)}{P_p(s(x_j^{t-1})|x^t)} < \beta$ **then**

add $s(x_{j+1}^{t-1})$ *to* $M(x^t)$;

$nc = C$;

delete all s *in* $\underline{M}(C, s(x_{j+1}^{t-1}))$ *from* $M(x^t)$;

end

2. 補題の証明

任意の $m_{s'}$ について,

$$\lim_{n\to\infty} \frac{P(x^n|m_{s'})}{P(x^n|m_s)} = 0 \quad a.s., \tag{A·1}$$

となる m_s が存在することを示す. 性質3.2より,

$$P(\theta^{\hat{m}}|m, x^n) = (\frac{n}{2\pi})^{\frac{k_m}{2}} \sqrt{\det I(\hat{\theta}|m)} + o(n^{\frac{k_m}{2}}), \tag{A·2}$$

である. また,

$$P(x^n|m) = \frac{P(x^n|m, \theta^{\hat{m}})P(\theta^{\hat{m}}|m)}{P(\theta^{\hat{m}}|m, x^n)}, \tag{A·3}$$

であるので, 任意の m_s と $m_{s'}$ に対して次の式が成り立つ.

$$\begin{aligned}\log \frac{P(x^n|m_{s'})}{P(x^n|m_s)} &= \log \frac{P(x^n|m_{s'}, \hat{\theta}^{m_{s'}})}{P(x^n|m_s, \hat{\theta}^{m_s})} + \frac{k_{m_s} - k_{m_{s'}}}{2} \log \frac{n}{2\pi} \\ &\quad - \log \frac{\sqrt{\det I(\theta^{*m_{s'}}|m_{s'})}}{\sqrt{\det I(\theta^{*m_s}|m_s)}} + \log \frac{P(\theta^{*m_{s'}}|m_{s'})}{P(\theta^{*m_s}|m_s)} \\ &\quad + o(1). \end{aligned} \tag{A·4}$$

ここで, $\log \frac{\sqrt{\det I(\theta^{*m_{s'}}|m_{s'})}}{\sqrt{\det I(\theta^{*m_s}|m_s)}}, \log \frac{P(\theta^{*m_{s'}}|m_{s'})}{P(\theta^{*m_s}|m_s)}$ は n に無関係な定数であることより以下を得る.

$$\begin{aligned}\log \frac{P(x^n|m_{s'})}{P(x^n|m_s)} &= \log \frac{P(x^n|m_{s'}, \hat{\theta}^{m_{s'}})}{P(x^n|m_s, \hat{\theta}^{m_s})} \\ &\quad + \frac{k_{m_s} - k_{m_{s'}}}{2} \log \frac{n}{2\pi} + O(1). \end{aligned} \tag{A·5}$$

次に $-\log P(x^n|m_s, \hat{\theta}^{m_s})$ を評価する.

まず, $-\log P(x^n|m_s, \hat{\theta}^{m_s})$ をテイラー展開すると,

$$\begin{aligned}&-\log P(x^n|m_s, \theta^{*m_s}) \\ &= -\log P(x^n|m_s, \hat{\theta}^{m_s}) \\ &\quad - \left.\frac{\partial \log P(x^n|m_s, \theta^{m_s})}{\partial \theta^{m_s}}\right|_{\theta^{m_s} = \hat{\theta}^{m_s}} (\theta^{*m_s} - \hat{\theta}^{m_s}) \\ &\quad + \frac{1}{2}(\theta^{*m_s} - \hat{\theta}^{m_s})^T \end{aligned} \tag{A·6}$$

$$\left.\frac{\partial^2 \log P(x^n|m_s,\theta^{m_s})}{\partial\theta^{m_s}{}^T\partial\theta^{m_s}}\right|_{\theta^{m_s}=\hat{\theta}^{m_s}}(\theta^{*m_s}-\hat{\theta}^{m_s})$$
$$+o((\theta^{*m_s}-\hat{\theta}^{m_s})^T$$
$$\left.\frac{\partial^2 \log P(x^n|m_s,\theta^{m_s})}{\partial\theta^{m_s}{}^T\partial\theta^{m_s}}\right|_{\theta^{m_s}=\hat{\theta}^{m_s}}(\theta^{*m_s}-\hat{\theta}^{m_s})).$$

ここで，$\left.\frac{\partial \log P(x^n|m_s,\theta^{m_s})}{\partial\theta^{m_s}}\right|_{\theta^{*m_s}=\hat{\theta}^{m_s}}=0$ である．また，大数の強法則より次の式が概収束の意味で成り立つ．

$$-\frac{1}{n}\left.\frac{\partial^2 \log P(x^n|m_s,\theta^{m_s})}{\partial\theta^{m_s}{}^T\partial\theta^{m_s}}\right|_{\theta^{m_s}=\hat{\theta}^{m_s}}$$
$$=I(\hat{\theta}^{m_s}|m_s)+o(1). \qquad \text{(A·7)}$$

つまり，$\frac{\partial^2 \log P(x^n|m_s,\theta^{m_s})}{\partial\theta^{m_s}{}^T\partial\theta^{m_s}}$ の部分は，Fisher 情報量の n 倍に概収束する．次に，$(\theta^{*m_s}-\hat{\theta}^{m_s})$ の部分を考える．系 3.1 より，

$$\sqrt{n}||\theta^{\hat{m}_s}-\theta^{*m_s}||=O\left((\log\log n)^{\frac{1}{2}}\right), \qquad \text{(A·8)}$$

であるので，結局以下を得る．

$$\frac{n}{2}(\theta^{*m_s}-\hat{\theta}^{m_s})^T I(\hat{\theta}^{m_s}|m_s)(\theta^{*m_s}-\hat{\theta}^{m_s})$$
$$=O(\log\log n)\ \ a.s. \qquad \text{(A·9)}$$

(A·9) 式を (A·6) 式に代入すると，

$$\log P(x^n|m_s,\hat{\theta}^{m_s}) \qquad \text{(A·10)}$$
$$=\log P(x^n|m_s,\theta^{*m_s})+O(\log\log n)\ \ a.s.$$

である．同様の事が $P(m_{s'}|x^n)$ にも成立するので結局，(17)(A·5)(A·10) 式から，

$$\log\frac{P(x^n|m_{s'})}{P(x^n|m_s)}$$
$$=\log\frac{P(x^n|m_{s'},\theta^{*m_{s'}})}{P(x^n|m_s,\theta^{*m_s})}$$
$$+O(\log\log n)+\frac{k_{m_s}-k_{m_{s'}}}{2}\log\frac{n}{2\pi}$$
$$=-nD(p^{\theta^*}_{m_{s'}};p^{\theta^*}_{m_s})+O(\log\log n)$$
$$+\frac{k_{m_s}-k_{m_{s'}}}{2}\log\frac{n}{2\pi}\ \ a.s., \qquad \text{(A·11)}$$

である．

ここで，$s^*\ll s\ll s'$ の場合を考える．明らかに任意の $m_{s'}$ に対して，以下を満たす m_s が存在する．

$$P(x^n|m_s,\theta^{*m_s})=P(x^n|m_{s'},\theta^{*m_{s'}}). \qquad \text{(A·12)}$$

上式を満たす m_s と $m_{s'}$ を考えると，$k_{m_{s'}}>k_{m_s}$ より，

$$\log\frac{P(x^n|m_{s'})}{P(x^n|m_s)}$$
$$=\log\frac{P(x^n|m_{s'},\theta^{*m_{s'}})}{P(x^n|m_s,\theta^{*m_s})}+\frac{k_{m_s}-k_{m_{s'}}}{2}\log\frac{n}{2\pi}$$
$$+O(\log\log n)$$
$$=0+\frac{k_{m_s}-k_{m_{s'}}}{2}\log\frac{n}{2\pi}+O(\log\log n)$$
$$\to-\infty\ \ a.s. \qquad \text{(A·13)}$$

となる．

一方，$s'\ll s\ll s^*$ の場合，$k_{m_{s'}}<k_{m_s}$ である．ここで，

$$\frac{1}{n}\log\frac{P(x^n|m_{s'},\theta^{*m_{s'}})}{P(x^n|m_s,\theta^{*m_s})}$$
$$=\frac{1}{n}\log\frac{\frac{P(x^n|m_{s^*},\theta^{*m_{s^*}})}{P(x^n|m_s,\theta^{*m_s})}}{\frac{P(x^n|m_{s^*},\theta^{*m_{s^*}})}{P(x^n|m_{s'},\theta^{*m_{s'}})}},$$
$$=\frac{1}{n}\log\frac{P(x^n|m_{s^*},\theta^{*m_{s^*}})}{P(x^n|m_s,\theta^{*m_s})}$$
$$-\frac{1}{n}\log\frac{P(x^n|m_{s^*},\theta^{*m_{s^*}})}{P(x^n|m_{s'},\theta^{*m_{s'}})}. \qquad \text{(A·14)}$$

が成立する．また仮定 3.1 より，

$$D(p^*_{m^*};p^*_{m_{s'}})-D(p^*_{m^*};p^*_{m_s})>0, \qquad \text{(A·15)}$$

であるので，(17) 式より，任意の $m_{s'}$ に対して

$$\log\frac{P(x^n|m_{s'},\theta^{*m_{s'}})}{P(x^n|m_s,\theta^{*m_s})}$$
$$=n\Big(D(p^*_{m^*};p^*_{m_s})-D(p^*_{m_{s^*}};p^*_{m_{s'}})\Big)+o(n)$$
$$\to-\infty\ \ a.s., \qquad \text{(A·16)}$$

となる m_s が存在することがわかる．

以上の議論より，

$$\log\frac{P(x^n|m_{s'})}{P(x^n|m_s)}$$
$$=\log\frac{P(x^n|m_{s'},\theta^{*m_{s'}})}{P(x^n|m_s,\theta^{*m_s})}+\frac{k_{m_{s'}}-k_{m_s}}{2}\log\frac{n}{2\pi}$$
$$+O(\log\log n)$$
$$=-O(n)+\frac{k_{m_{s'}}-k_{m_s}}{2}\log\frac{n}{2\pi}$$
$$+O(\log\log n)$$
$$\to-\infty\ \ a.s. \qquad \text{(A·17)}$$

を得る．$-O(n)$ は負の数を表す．

上式より任意の $m_{s'}$ に対して次式を満たす m_s が存在することがわかる．

$$\lim_{n\to\infty} \frac{P(x^n|m_{s'})}{P(x^n|m_s)} = 0 \quad a.s., \qquad \text{(A·18)}$$

故に補題は証明された． □

(平成年月日受付，月日再受付)

野村　亮　(正員)

平 8 早大・理工・工業経営卒．平 10 同大大学院修士課程了．平 11 同大博士後期課程入学．平 12 同大助手．情報源符号化に関する研究に従事．IEEE，情報理論とその応用学会，日本経営工学会等各会員．

松嶋　敏泰　(正員)

昭 53 早大・理工・工業経営卒．昭 55 同大大学院修士課程了．同年，日本電気 (株) 入社．昭 61 早大・理工学研究科・博士後期課程入学．平 1 横浜商科大学講師．平 3 同大助教授．平 4 早大・理工学部・工業経営学科 (現在経営システム工学科) 助教授，平 9 同大教授，現在に至る．知識情報処理および情報理論とその応用に関する研究に従事．工学博士．平 13 ハワイ大学客員研究員．IEEE，情報理論とその応用学会，人工知能学会，情報処理学会，OR 学会，日本経営工学会等各会員．

平澤　茂一　(正員)

昭 36 早大・理工・数学卒．昭 38 同電気通信卒．同年三菱電機 (株) 入社．昭 56 早大・理工・工業経営学科 (現在経営システム工学科) 教授，現在に至る．情報理論とその応用，データ伝送方式，ならびに計算機応用システムの開発などの研究に従事．工学博士．昭 54UCLA 計算機科学科客員研究員．昭 60 ハンガリー科学アカデミー，昭 61 伊トリエステ大学客員研究員．平 5 電子情報通信学会 小林記念特別賞，業績賞受賞．IEEE Fellow，情報理論とその応用学会，人工知能学会，情報処理学会，OR 学会，日本経営工学会等各会員．

On the variance and the probability of length overflow of lossless codes

Ryo NOMURA[1]
Waseda University
Shinjuku-ku, Tokyo, Japan.
ryochan@matsu.mgmt.waseda.ac.jp

Toshiyasu MATSUSHIMA
Waseda University
Shinjuku-ku, Tokyo, Japan.

Shigeichi HIRASAWA
Waseda University
Shinjuku-ku, Tokyo, Japan.

Abstract — **In this paper, we show the probability of length overflow of several codes by using the variance and the asymptotic normality of the codelength.**

I. Introduction

Lossless source coding schemes are examined under several criterions. The most representative criterion is redundancy. Recently, Merhav[1] proposed the probability of length overflow.

In this paper we redefine the probability of length overflow. We consider a finite alphabet source $A = \{i : 0 \le i \le k-1\}$. Let $x^n = x_1x_2x_3\cdots x_n \in X^n$ denotes a source sequence. And let $p(x^n)$ denotes the probability distribution of a source. Let $L(\cdot)$ be a codelength and ϵ_n be a function of n.

Definition I. 1 *The probability of length overflow is defined by*

$$Pr\{L(x^n) > \epsilon_n\}. \tag{1}$$

We shall evaluate a code by using the probability of length overflow instead of the expected codelength.

Next we define the two quantities, that have very important role in this paper. First we generalize the minimal coding variance, which is inherent value of a source, proposed by Kontoyiannis[2].

Definition I. 2 *The rth moment of self-information is defined by*

$$M(X)^r = \lim_{n\to\infty} E\left[\left\{-\frac{1}{n}\log p(x^n) - E[-\frac{1}{n}\log p(X^n)]\right\}^r\right].$$

Especially, the 2nd moment of self-information coincides with the minimal coding variance.

Second we define the moment of codelength.

Definition I. 3 *Let $L_c(x^n)$ denotes the codelength for sequence x^n when we use a code c. Then the rth moment of a code c is denoted by*

$$\sigma_c^r = \lim_{n\to\infty} E\left[\left\{\frac{1}{n}L_c(x^n) - E[\frac{1}{n}L_c(X^n)]\right\}^r\right]. \tag{2}$$

Especially, when $r = 2$ we call this the variance of codelength of a code c.

II. The probability of length overflow

We show the probability of length overflow of a code c. Let $L_c(x^n)$ denote the codelength of a code c for x^n.

Lemma II. 1 *If the codelength of a code c satisfies asymptotic normality with respect to a source, the probability of length overflow of a code c is given by*

$$\lim_{n\to\infty} Pr\{L_c(x^n) \ge \epsilon_n\} \tag{3}$$
$$= \int_{Z_n^*}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{y^2}{2}\right]dy,$$

where, $Z_n^* = \frac{\epsilon_n - E[L_c(x^n)]}{\sqrt{n\sigma_c^2}}$, σ_c^2 is the variance of a code c.

[1]This work was supported by in part of Waseda University under Grant 99A-551 for Special Research Projects.

When a source distribution is known, it is well known that a codelength $-\log p(x^n)$ minimize the expected codelength. We call this code Shannon code and let σ_S^2 be the variance of Shannon code. Obviously, the variance of Shannon codelength coincides with 2nd moment of self-information. Here we define a condition of a source as follows.

Condition II. 1 *The codelength of Shannon code with respect to a source satisfies the asymptotic normality.*

Then we have the following lemma.

Lemma II. 2 *Under ConditionII. 1, if* $\lim \epsilon_n > nH(X) + \sqrt{nM(X)^2}$, *then we have*

$$\lim_{n\to\infty} Pr\{-\log p(x^n) > \epsilon_n\} = 0. \tag{4}$$

III. The probability of length overflow of Bayes code

We consider a parameterized source distribution. Let $\theta \in \Theta$ is a k-dimensional parameter of a source. If θ is unknown, it is known that Bayes code minimize the redundancy with respect to Bayes criterion. The coding probability of Bayes code is given by $m(x^n) = \int_{\theta\in\Theta} p(x^n|\theta)p(\theta)d\theta$, where $p(\theta)$ is a prior distribution of θ. We define a condition of a source.

Condition III. 1 *The codelength of Bayes code with respect to a source satisfies the asymptotic normality.*

Then we have the following theorem.

Theorem III. 1 *Let the variance of Bayes code denoted by σ_B^2, we have*

$$M(X)^2 + \frac{k}{n} \ge \sigma_B^2 \ge M(X)^2 + \frac{k}{n} - \sqrt{\frac{2kM(X)^2}{n}}. \tag{5}$$

From above theorem, we have the following lemma.

Lemma III. 1 *Under ConditionIII. 1, if* $\lim_{n\to\infty}\epsilon_n > nH(X) + \sqrt{nM(X)^2}$, *then we have*

$$\lim_{n\to\infty} Pr\{-\log m(x^n) > \epsilon_n\} = 0. \tag{6}$$

IV. Consideration

We obtained the probability of length overflow of codes, that minimize the expected codelength. From above lemmas neither source distribution is known or unknown, under ConditionII. 1 ,III. 1 , if we wish the probability of length overflow goes to 0 then it is necessary that ,$\lim_{n\to\infty}\epsilon_n > nH(X) + \sqrt{nM(X)^2}$.

We introduce the moment of self-information and the moment of codelength, that play very important role to analyse the probability of length overflow.

References

[1] N. Merhav, "Universal Coding with Minimum Probability of Codeword Length Overflow," *IEEE Trans. Inf. Theory*,37(3):556-563,1991.

[2] I. Kontoyiannis, "Second-Order Noiseless Source Coding Theorems," *IEEE Trans. Inf. Theory*,43(4):1339-1341,1997.

0-7803-5857-0/00/$10.00 ©2000 IEEE.

Achievable rates of random number generators for an arbitrary prescribed distribution from an arbitrary given distribution

Takahiro Yoshida*
School of Science and Engneering
Waseda University
Shinjuku-ku, Tokyo, Japan.
takahiro@matsu.mgmt.waseda.ac.jp

Toshiyasu Matsushima
School of Science and Engneering
Waseda University
Shinjuku-ku, Tokyo, Japan.

Shigeichi Hirasawa
School of Science and Engneering
Waseda University
Shinjuku-ku, Tokyo, Japan.

Abstract — **In this paper, we show maximal rates in the case that random number generators generate a random sequence with an arbitrary prescribed distribution from a random sequence with an arbitrary given distribution.**

I. Introduction

One of generalizing the random number generation problem is to relax the requirement that the target random numbers should be generated exactly according to the prescribed distribution. We are especially concerned with the case of the fixed length random number generation. Let $\mathcal{X}$ and $\mathcal{Y}$ be countable infinite set. Let us define a general source as an infinite sequence $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ of n-dimensional random variable X^n taking value in $\mathcal{X}^n$ and $\mathbf{Y} = \{Y^m\}_{m=1}^{\infty}$ of m-dimensional random variable Y^m taking value in $\mathcal{Y}^m$.

In this paper, we shall investigate into maximal rate in the case that random number generators generate a random sequence with an arbitrary prescribed distribution from a random sequence with an arbitrary given distribution in the sense of vanishing variational distance. The variational distance between two distributions P_z and $P_{\hat{z}}$ on $\mathcal{Z}$ is defined as follows

$$d\left(Z,\hat{Z}\right) = \sum_{z\in\mathcal{Z}} |P_Z(z) - P_{\hat{Z}}(z)|. \tag{1}$$

In this setting, there are two types of the case for the fixed length random number generation. One is that every source n_m symbol realization is deterministically transformed into a sequence with length m where n_m depends only on m. The other is that every source n symbol realization is deterministically transformed into a sequence with length m_n.

II. Formulation of the problem

Definition II. 1 *R is called a type A achievable rate for the source $\mathbf{X}$ and $\mathbf{Y}$ if there exists a sequence of mappings $\varphi_n : \mathcal{X}^n \to \mathcal{Y}^{m_n}$ such that*

$$\liminf_{n\to\infty} \frac{m_n}{n} \geq R \tag{2}$$

and

$$\lim_{n\to\infty} d\left(Y^{m_n}, \varphi_n(X^n)\right) = 0. \tag{3}$$

Moreover the supremum of R that are type A achievable rate for the source $\mathbf{X}$ and $\mathbf{Y}$ is denoted by $S_A(\mathbf{X},\mathbf{Y})$ which we call maximal type A achievable rate.

Definition II. 2 *R is called a type B achievable rate for the source $\mathbf{X}$ and $\mathbf{Y}$ if there exists a sequence of mappings $\varphi_m : \mathcal{X}^{n_m} \to \mathcal{Y}^m$ satisfying the condition that n_m and m replace n and m_n respectively in Formula (2) and (3). Moreover the supremum of R that are type B achievable rate for the source $\mathbf{X}$ and $\mathbf{Y}$ is denoted by $S_B(\mathbf{X},\mathbf{Y})$ which we call maximal type B achievable rate.*

*This research was supported in part of Waseda University under Grant 99A-551 for Special Research Projects.

III. Main Results

We denote the *limsup in probability* of $\left\{\frac{1}{n}\log\frac{1}{P_{Z^n}(Z^n)}\right\}_{n=1}^{\infty}$ and the *liminf in probability* of that by $\overline{H}(\mathbf{Z})$ and $\underline{H}(\mathbf{Z})$, respectively[1][2][3]. Then we have

Theorem III. 1

$$\frac{\underline{H}(\mathbf{X})}{\overline{H}(\mathbf{Y})} \leq S_A(\mathbf{X},\mathbf{Y}) \leq \min\left(\frac{\underline{H}(\mathbf{X})}{\underline{H}(\mathbf{Y})}, \frac{\overline{H}(\mathbf{X})}{\overline{H}(\mathbf{Y})}\right), \tag{4}$$

$$\frac{\underline{H}(\mathbf{X})}{\overline{H}(\mathbf{Y})} \leq S_B(\mathbf{X},\mathbf{Y}) \leq \min\left(\frac{\underline{H}(\mathbf{X})}{\underline{H}(\mathbf{Y})}, \frac{\overline{H}(\mathbf{X})}{\overline{H}(\mathbf{Y})}\right). \tag{5}$$

We notice that if either source $\mathbf{X}$ or source $\mathbf{Y}$ satisfies the *strong converse property*[3], then

$$S_A(\mathbf{X},\mathbf{Y}) = S_B(\mathbf{X},\mathbf{Y}) = \frac{\underline{H}(\mathbf{X})}{\overline{H}(\mathbf{Y})}. \tag{6}$$

In the case that source $\mathbf{Y}$ is uniform distribution, i.e., $P_Y(Y) = 1/M$ $(M < \infty)$, by replacing m_n with $\log M_n$ in Formula (2) of definition II. 1 , it is equivalent to the intrinsic randomness problem defined by Vembu and Verdú[1]. Then,

$$S_A(\mathbf{X},\mathbf{Y}) = \underline{H}(\mathbf{X}), \tag{7}$$

where $M_n = M^{m_n}$. On the other hand, in the case that source $\mathbf{X}$ is uniform distribution, i.e., $P_X(X) = 1/M$, by replacing n_m with $\log M_m$ in definition II. 2 , the minimum of reciprocal number of type B achievable rate is equivalent to the minimal achievable resolvability rate defined by Han and Verdú[2], i.e.,

$$\frac{1}{S_B(\mathbf{X},\mathbf{Y})} = \overline{H}(\mathbf{Y}), \tag{8}$$

where $M_m = M^{n_m}$.

For the reasons stated above, essence of which maximal achievable rate is uniquely decidable is that either source $\mathbf{X}$ or source $\mathbf{Y}$ satisfies the strong converse property. Since uniform distribution satisfy the strong converse property, Both maximal achievable intrinsic randomness rate[1] and minimal achievable resolvability rate[2] are the special case of theorem III. 1 .

IV. Conclusion

We have defined two types of random number generation problem and obtained two maximal achievable rates. Both intrinsic randomness problem[1] and resolvability problem[2] are the special case of our result.

References

[1] S. Vembu and S. Verdú, "Generating Random Bits from an Arbitrary Source: Fundamental Limits," *IEEE Trans. Inf. Theory*, vol. 41, no.5, pp.1322-1332, Sept. 1995

[2] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inf. Theory*, vol.39, no.3, pp.752-772, May 1993

[3] T. S. Han, "Information-Spectrum methods in information theory," Baifukan, Tokyo, 1998 (In Japanese)

0-7803-5857-0/00/$10.00 ©2000 IEEE.

On Analysis of noiseless decision feedback scheme using fixed size list decoder for tree codes

Toshihiro Niinomi
Kanagawa Institute of Tech.,
1030 Shimo-Ogino, Atsugi-shi,
Kanagawa, 243-0292 Japan
e-mail:
niinomi@ele.kanagawa-it.ac.jp

Toshiyasu Matsushima
School of Science and Engineering,
Waseda University,
3-4-1 Ohkubo, Shinjyuku-ku,
Tokyo,169-0072 Japan

Shigeichi Hirasawa
School of Science and Engineering,
Waseda University,
3-4-1 Ohkubo, Shinjyuku-ku,
Tokyo,169-0072 Japan

Abstract — **In this paper, the generalized proof are shown for the coding theorem of [1]. Consequently, the further discussion is obtained.**

I. Introduction

Despite list decoder is seldom employed for applications, Generalized Viterbi Algorithm(GVA) use list decoder for its mean process [2]. In the analysis of the GVA, the error probability is dominated by that of list decoder. Therefore, we proposed the decision feedback scheme with GVA, which has the constructive decision rule for list decoding with feedback [1].

In this paper, the more strict bound and a certain generalization of the proof of [1] are shown. As a result, the further properties can be clarified. Throughout this paper, DMC characterized by $P = \{P_{ij},\ j \in A, i \in B\}$, and noiseless feedback is assumed.

II. The algorithm [1]

We are concerned with a q-ary tree code, each branch of which is assigned with v input alphabet. So, the rate of the code is defined by $R = \frac{1}{v}\ln q$ .A N information sequence of q-ary alphabet specifies a path, denoted by $\mathbf{u}_i^N$. Let the subsequence of path $\mathbf{u}_i^N$ from the root to the n-th level be $\mathbf{u}_i^n$. Furthermore, L(branches) is defined as decoding constraint length of the GVA.

{ *Initial condition and Recursive procedure* }

(Step 1)*Initial condition* : At the level $n-1$, each state of q^{L-1} has their list, namely S survivors. Each survivor of the list is labeled "Accept".

For the level n $(L \le n \le N)$, repeat (Step 2)~(Step 4), recursively.

(Step 2)*Path extension*: At the level n, all retained paths are extended by one branch as $\mathbf{u}^n = \mathbf{u}^{n-1}u$. Then each metric of Sq^L paths is calculated.

(Step 3)*Path selection*: At each state of q^{L-1}, the best S paths are selected among the qS paths as the list.

(Step 4)*Testing*: We denote the selected list as $\mathcal{L}$, $\mathcal{L} = \{\mathbf{u}_{(1)}^n, \mathbf{u}_{(2)}^n, \cdots, \mathbf{u}_{(S)}^n\}$, and $\mathbf{u}_{(S+1)}^n$ is the $S+1$-th most path at the state [1]. The listed paths are labeled "Accept" or "Reject" by the following decision rule. However, a path once labeled "Reject" is kept its label "Reject". The rule is if $\frac{Pr(\mathbf{y}^{vn}|\mathbf{u}_{(1)}{}^n)}{Pr(\mathbf{y}^{vn}|\mathbf{u}_{(S+1)}{}^n)} \le \Delta, \Delta \ge 1$ holds, $\mathbf{u}_{(i)}^n, i = 1,2,\cdots,S$ are labeled "Reject". Otherwise, $\mathbf{u}_{(i)}^n, i = 1,2,\cdots,S$ are labeled "Accept". If there is no survivor labeled "Accept" at any q^{L-1} state, the retransmission is required and restart from Step 1.

{ *Final path selection at the check tail* }

By $L-1$ known symbols, q^{L-1} lists are reduced to one list with Step 2 ~Step 4. Then, by $T-(L-1)$ known symbols, the best path is selected among the S survivors of the final list. If the label of the best path is "Reject", the retransmission is required and restart from Step 1.

[1]For the received sub-sequence $\mathbf{y}^{vn}$ from root to level n , we denote the likelihood of the k-th most path as $Pr(\mathbf{y}^{vn}|\mathbf{u}_{(k)}{}^n)$.

III. Main results

Though the analysis in [1] depends on each tree configuration [5], the bounds newly obtained are independent. So, we newly observe the case that S is very large. For obtaining the feedback exponent $-\frac{1}{vL}\ln Pr(E_2)$ of Forney [4], we take Δ as $-\frac{1}{vL}\ln Pr(E_1) \to 0$ $(L \to \infty)$, where $Pr(E_1)$ and $Pr(E_2)$ is Pr[The decoding error occurs, or, the retransmission is required.] and Pr[The decoding error occurs.], respectively. We show this result as Theorem.

[Theorem]As $S \to \infty$, the exponent approaches to $e_1^{(\infty)}(R)$,

$$e_1^{(\infty)}(R) = \max_{\mathbf{q},\alpha,\beta \in \mathcal{D}_4} \left\{ E_o(1,\alpha,\beta,\mathbf{q}) + \alpha \cdot e_F^{(1)}(R) \right\},$$

$$\mathcal{D}_4 = \{\alpha \ge 0,\ \beta \ge 0, \epsilon_e = E_o(1,\alpha,\beta,\mathbf{q}) - \beta R > 0\}.$$

$$E_o(S,\sigma_x,\rho_x,\mathbf{q})$$

$$= -\ln\left[\sum_{j\in B}\left(\sum_{i\in A} q_i P_{ji}^{1-S\sigma_x}\right)\left(\sum_{k\in A} q_k P_{jk}^{\sigma_x/\rho_x}\right)^{S\rho_x}\right],$$

$$e_F^{(S)}(R) = \max_{\mathbf{q},\nu\in\mathcal{D}_3} E_{oF}(S,\nu,\mathbf{q}),$$

$$\mathcal{D}_3 = \{E_{oF}(S,\nu,\mathbf{q}) - \nu SR > 0,\quad \nu > 0\}$$

$$E_{oF}(S,\nu,\mathbf{q}) = S\sum_{k\in B}\sum_{j\in A} q_j P_{kj} \ln\left[\frac{P_{kj}^{1/\nu}}{\sum_{j\in A} q_j P_{kj}{}^{1/\nu}}\right]^{\nu}$$

IV. Conclusion

We show the properties of the decision feedback scheme using fixed size list decoder, especially the size of list is very large. The exponents we have obtained have the similar properties to those of the GVA.

References

[1] T.Niinomi, T.Matsushima and S.Hirasawa, "A decision feedback scheme using list decoding for tree codes", IEICE Trans. A, Vol.83, No.1, Jan. 2000 (in Japanese).

[2] T.Hashimoto, "A list-type reduced-constraint generalization of the Viterbi algorithm", IEEE Trans. Inf.Theory vol.IT-33, pp.866-876, Nov.1987.

[3] T.Hashimoto, "On the error exponent of convolutionaly coded ARQ", IEEE Trans. Inf.Theory vol.IT-40, pp.567-575, Mar.1994.

[4] G.D.Forney,Jr., "Exponential error bounds for erasure, list and decision feedback schemes", IEEE Trans.Inf.Theory vol.IT-14, pp.206-220, Mar.1968.

[5] G.D.Forney,Jr., "Convolutional codes II : Maximum likelihood decoding", Inf.Control. vol.25, pp.222-266, Jul.1974.

0-7803-5857-0/00/$10.00 ©2000 IEEE.

The 23rd Symposium on Information Theory
and Its Applications (SITA2000)
Aso, Kumamoto, Japan, Oct. 10 - 13, 2000

不確実性を含む演繹推論に関する一考察
On a Deductive Reasoning with Uncertainty

鈴木 誠* 松嶋 敏泰* 平澤 茂一*
Makoto SUZUKI Toshiyasu MATSUSHIMA Shigeichi HIRASAWA

Abstract— In this paper, we shall discuss a problem of probabilistic reasoning. We shall divide the reasoning into two parts, i.e. deduction and induction. The induction is a calculation of the maximum likelihood estimator of each cell of a contingency table. On the other hand, the deduction is a calculation of conditional probabilities using the maximum likelihood estimator when several marginal sums of the conditional probabilities are given. We propose a new reasoning method that guarantees a given reliability based on statistics by applying the interval estimation.

Keywords—contingency table, Fisher information matrix, information inequality, interval estimation

1 はじめに

1980年代にBelief Network(BN)がPearlによって提案されて以来，BNは確率的推論の分野の主要な手法として活発に研究されている[2]．BNとは，事象(命題)をノード，事象間の確率的依存関係をアークで表現した有向非循環グラフ(Directed Acyclic Graph : DAG)を用いて確率的な推論を実現する手法である．PearlのBNは，事象間の因果関係等の不確実性を含む知識が条件付確率表(Conditional Probability Table : CPT)で与えられることを前提として，それらの事象間の関係をDAGで表現し有効な確率的推論法を提案している．

前稿では，Pearlの研究においてCPTが与えられることと同様の立場にたち，真の確率分布を規定する母数が既知であるという仮定のもとで，不確実な知識を用いた演繹推論をモデル化し，その推論が“ある確率変数に対する広義の条件付確率の計算”であることを明らかにした[3]．そして，その条件付確率の計算を実現する演繹推論アルゴリズムとして，離散データ解析の分野で用いられているISP(Iterative Scaling Procedure)[1]を利用した反復アルゴリズムを提案した．しかし，実用的な見地に立つと真の母数が既知であるという仮定は必ずしも適当であるとは言えず，母数をどのように獲得するかという問題点が残されていた．

そこで本稿では統計学的な見地にたち，既知情報が真の母数ではなく分割表形式のデータである場合の不確実性を含む推論の問題を論じる．したがって本稿で論じる推論処理は，分割表形式のデータから母数を推定し，その母数の推定値と一つの個体についての観測事実を用いて未観測の確率変数についての条件付確率を計算するという二段階の推論プロセスに分類される．以降では，前者のデータから母数を推定する推論プロセスを帰納推論と呼び，後者の条件付確率を計算する推論プロセスを演繹推論と呼ぶこととする．

そして本稿の目的は，上記の推論モデルのもとで，区間推定の理論を応用することにより確信度に幅を持たせて推論結果の信頼性を保証する推論法を提案することである．この推論結果は統計学的には最尤推定量に基づく信頼区間となっており，漸近的かつ近似的に区間の幅を最小にしている．

* 早稲田大学理工学部経営システム工学科，〒169-8555 東京都新宿区大久保3-4-1. Department of Industrial and Management Systems Engineering, School of Science and Engineering, Waseda University, 3-4-1, Ohkubo, Shinjuku, Tokyo, 169, Japan. makoto@hirasa.mgmt.waseda.ac.jp

2 準備

2.1 表記法の整理

本稿で用いる記号を以下にまとめる†.

k	:	基本式の数
s	:	観測が行われた基本式の数(観測事実の数)
X_i	:	基本式$A_i(\omega)$の真理値，確率変数 $X_i \in \{1,0\}$, $(i=1,2,\cdots,k)$ X_iの実現値は小文字x_iで表記する．
$\mathbf{d}^{(\ell)}$	:	l番目のサンプル，2^k次元単位ベクトル $\mathbf{d}^{(\ell)} = (d_1{}^{(\ell)}, d_2{}^{(\ell)}, \cdots, d_{2^k}{}^{(\ell)})'$
$\mathbf{c}$	:	分割表のセルの頻度，2^k次元列ベクトル $\mathbf{c} = (c_1, c_2, \cdots, c_{2^k})'$, $c_i = \sum_{\ell=1}^{N} d_i{}^{(\ell)}$
Ξ	:	真の(多項)分布を規定する母数(真の母数) 2^k次元列ベクトル $\Xi = (\xi_1, \xi_2, \cdots, \xi_{2^k})'$ 以降はd(=domain)結合確信度と呼ぶ． 式(24),(26)の$P(X_1, \cdots, X_k \mid \Xi)$は$\xi_{x_1 x_2 x_3}$
$\hat{\Xi}(\mathbf{c})$	:	真の(多項)分布を規定する母数の推定値 以降は$\hat{\Xi}$と表記する． 2^k次元列ベクトル $\hat{\Xi} = (\hat{\xi}_1, \hat{\xi}_2, \cdots, \hat{\xi}_{2^k})'$
$\mathbf{q}$	:	s個の基本式$A_{r_m}(\omega)$に関する観測事実 $(m=1,2,\cdots,s,\ \ r_m \in \{1,\cdots,k\})$ s次元列ベクトル $\mathbf{q} = (q_{r_1}, q_{r_2}, \cdots, q_{r_s})'$ 各q_{r_m}は式(25)のi(=individual)確信度
$\mathbf{y}$	:	観測過程によって得られた基本式$A_{r_m}(\omega)$の真理値，通信路を経て得られた受信語 s次元列ベクトル $\mathbf{y} = (y_{r_1}, y_{r_2}, \cdots, y_{r_s})'$
Ψ	:	観測過程(通信路)を規定するパラメータ s次元列ベクトル $\Psi = (\psi_{r_1}, \psi_{r_2}, \cdots, \psi_{r_s})'$ 式(24),(26)の$P(y_i \mid X_i, \psi_i)$はψ_{r_m}
$\mathbf{p}(\Xi, \mathbf{q})$	:	真の母数Ξと観測事実$\mathbf{q}$が得られたもとでの条件付確率の(真の)値 式(26)(または,(16))のi結合確信度 以降は単に$\mathbf{p}$と表記する． 2^k次元列ベクトル$\mathbf{p} = (p_1, p_2, \cdots, p_{2^k})'$
$\mathbf{p}(\hat{\Xi}, \mathbf{q})$	:	母数の推定値$\hat{\Xi}$と観測事実$\mathbf{q}$が得られたもとでの条件付確率の値，以降は$\hat{\mathbf{p}}$と表記する．2^k次元列ベクトル$\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_{2^k})'$

2.2 2つの不確実性と推論システムの入出力

前稿では不確実性を含む演繹推論をモデル化した．その際，演繹推論で扱っている不確実性を“(1)ドメイン全体に対する不確実性”と“(2)個体に対する不確実性”の2つに分類した．本稿では，不確実性を含む推論を帰納推論と演繹推論の二段階に分けて考察し，前稿と同様に(1)，(2)の不確実性を分類するという視点から帰納・演繹の各推論を整理する．推論システム全体の入出力，各推論モジュールの入出力とそれらが扱う上記(1)と(2)の不確実性の関係を図1に示す．

† ξ_iと$\xi_{x_1 \cdots x_k}$，またはp_iと$p_{x_1 \cdots x_k}$は添え字の振り方が10進表現と2進表現の違いはあるが全く同じものを意味する．$\xi_{x_1 \cdots x_k}$と$p_{x_1 \cdots x_k}$は基本式との対応を分かりやすくするための表現である．

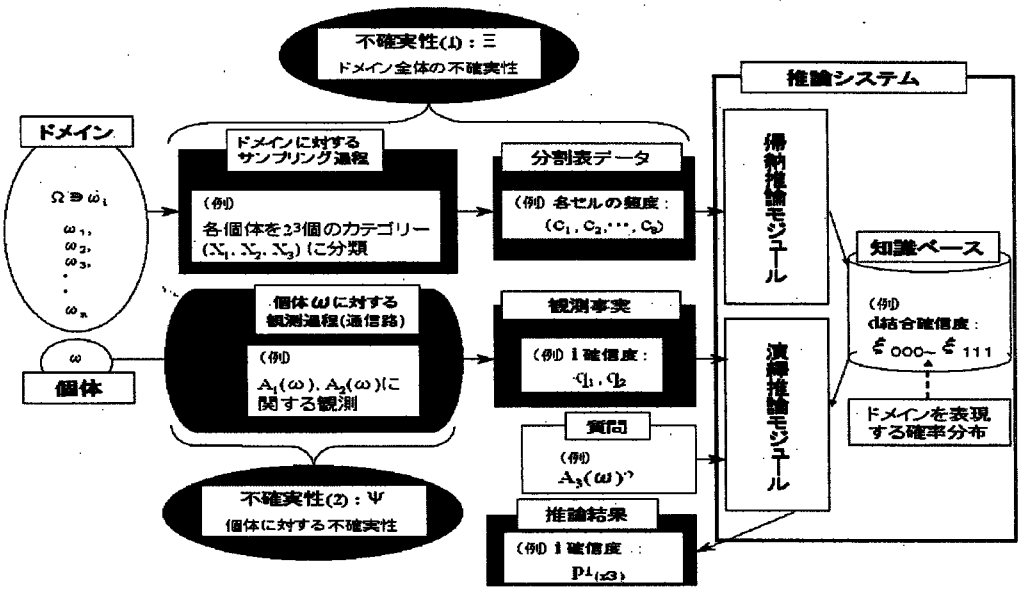

図 1: 2つの不確実性と推論システムの入出力

帰納推論モジュールはドメイン全体に対するサンプリングによって得られた分割表データ $\mathbf{c}$ を入力とし，母数の推定値 $\hat{\Xi}$ を出力する．すなわち帰納推論は，従来の統計学で通常行われている推定であり，ドメイン全体に対する不確実性のみを扱っている．

これに対し演繹推論モジュールは，(a) 帰納推論による母数の推定値 $\hat{\Xi}$ と，(b) いくつかの基本式についての観測事実 $\mathbf{q}$ を入力とし，未観測の基本式についての確信度 $\mathbf{p}(\hat{\Xi}, \mathbf{q})$ (厳密には $\mathbf{p}(\hat{\Xi}, \mathbf{q})$ の周辺和) を出力する．ここで，前者 (a) の入力はドメイン全体に対する不確実性を含んでいるのに対し，後者 (b) の入力は個体に対する不確実性を意味している点が重要である．このように，演繹推論では両方の不確実性を扱っている．

2.3 前稿との違い

前稿との立場の大きな違いは，演繹推論の入力情報の一つであるドメイン全体の不確実性を表現する真の確率分布が既知か未知かという点である．つまり，前稿では真の分布を規定する母数が既知であると仮定していた．したがって，帰納推論が必要とされていなかった．これに対し，本稿ではその母数が未知であると仮定し，分割表形式のデータをもとに帰納推論によって推定された値を演繹推論の入力としている．実用性を考慮すると，専門家が真の分布を与えることは困難であり，データからそれを推定しようという本稿の立場は有用であると言える．

3 推論結果の信頼性を保証する演繹推論法

3.1 演繹推論結果における推定誤差

演繹推論における条件付確率 $\hat{\mathbf{p}}$ に推定誤差が生じる原因を定性的に考えると以下のようになる．帰納推論では分割表形式のデータから母数の最尤推定を行うので推定誤差が混入する．一方，演繹推論では単なる条件付確率の計算を行っているだけであるので，推定誤差が入り込む余地はない．しかし，その条件付確率計算で用いる最尤推定量 $\hat{\Xi}$ が推定誤差を含んでいるため，計算された条件付確率 $\hat{\mathbf{p}}$ も推定誤差を含むことになる．別の見方をすれば，演繹推論は式 (26)(または，(16)) で示されるように $\hat{\Xi}$ から $\hat{\mathbf{p}}$ への一種のパラメータ変換を行っているとみなすこともできる．

3.2 最尤推定量とクラメール・ラオの不等式

最尤推定量の漸近有効性により，サンプル数 N が大きいとき，最尤推定量 $\hat{\Xi}$ は不偏推定量であり，$\hat{\Xi}$ の分散 $Var(\hat{\Xi})$ は多母数の場合のクラメール・ラオの不等式の下限を達成する．すなわち，次式が成り立つ．

$$Var(\hat{\Xi}) \geq J(\Xi) I(\Xi)^{-1} J(\Xi)' \tag{1}$$

ここで，$I(\Xi)$ は $(2^k-1) \times (2^k-1)$ のフィッシャー情報行列‡で，$I(\Xi)^{-1}$ はその逆行列である．また，$J(\Xi)$ は $(2^k-1) \times (2^k-1)$ のヤコビ行列で，$J(\Xi)'$ はその転置行列である．したがって，$\hat{\mathbf{p}}$ の分散の最小値は，次節で述べる $\hat{\Xi}$ から $\hat{\mathbf{p}}$ へのパラメータ変換によって引き起こされるフィッシャー情報行列の変換を考えることにより得られる．

3.3 フィッシャー情報行列の変換

ここでは，式 (26)(または，(16)) の "$\hat{\Xi}$ から $\hat{\mathbf{p}}$ へのパラメータ変換" によって引き起こされるフィッシャー情報行列の変換について論じる [4]．

以降では，対数尤度関数を $l(\Xi(\mathbf{p}), \mathbf{c})$ (または単に l) と表記する．ここで，式 (26)(または，(16)) によれば $\Xi = \Xi(\mathbf{p})$，すなわち $\xi_i = \xi_i(p_1, p_2, \cdots, p_{2^k})$ という関係が成り立っていることに注意されたい．

ところで，合成関数の微分の鎖則により次式が成り立つ．

$$\frac{\partial l(\Xi(\mathbf{p}), \mathbf{c})}{\partial p_i} = \sum_{j=1}^{2^k-1} \frac{\partial \xi_j}{\partial p_i} \frac{\partial l(\Xi(\mathbf{p}), \mathbf{c})}{\partial \xi_j} \tag{2}$$

これを行列表現すると次式のようになる．

$$\begin{bmatrix} \frac{\partial l}{\partial p_1} \\ \cdot \\ \cdot \\ \frac{\partial l}{\partial p_{2^k-1}} \end{bmatrix} = \begin{bmatrix} \frac{\partial \xi_1}{\partial p_1} & \cdots & \frac{\partial \xi_{2^k-1}}{\partial p_1} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \frac{\partial \xi_1}{\partial p_{2^k-1}} & \cdots & \frac{\partial \xi_{2^k-1}}{\partial p_{2^k-1}} \end{bmatrix} \begin{bmatrix} \frac{\partial l}{\partial \xi_1} \\ \cdot \\ \cdot \\ \frac{\partial l}{\partial \xi_{2^k-1}} \end{bmatrix} \tag{3}$$

‡ 多項分布の母数 Ξ に含まれる全ての ξ_i の和は1となる，すなわち $\sum_{i=1}^{2^k} \xi_i = 1$ であるので母数の自由度は 2^k-1 であり，フィッシャー情報行列 $I(\Xi)$ は $(2^k-1) \times (2^k-1)$ 行列となる．

$\Xi=\Xi(\mathbf{p})$の$(2^k-1)\times(2^k-1)$のヤコビ行列$J(\frac{\partial\xi}{\partial p})$の$(i,j)$要素を$[\frac{\partial\xi_i}{\partial p_j}]$とおき，対数尤度$l(\Xi(\mathbf{p}),\mathbf{c})$を$\mathbf{p}$および$\Xi$の要素で偏微分して得られる列ベクトルをそれぞれ$[\frac{\partial l}{\partial p_j}]$, $[\frac{\partial l}{\partial \xi_i}]$で表せば，式(3)は次式で表すことができる．

$$[\frac{\partial l}{\partial p_j}]=J(\frac{\partial\xi}{\partial p})'[\frac{\partial l}{\partial\xi_i}] \tag{4}$$

式(4)の両辺の分散共分散行列を計算すれば$\mathbf{p}$およびΞに関するフィッシャー情報行列の間には次式で表される関係があることがわかる．

$$I(\mathbf{p})=J(\frac{\partial\xi}{\partial p})'I(\Xi)J(\frac{\partial\xi}{\partial p}) \tag{5}$$

今，$\mathbf{p}$とΞが1対1の関係にあり，変換のヤコビ行列は正則であるので，式(5)の逆行列を求めることにより以下のようになる．

$$\begin{aligned} I(\mathbf{p})^{-1} &= J(\frac{\partial\xi}{\partial p})^{-1}I(\Xi)^{-1}J(\frac{\partial\xi}{\partial p})'^{-1} \\ &= J(\frac{\partial p}{\partial\xi})I(\Xi)^{-1}J(\frac{\partial p}{\partial\xi})' \end{aligned} \tag{6}$$

式(6)では，$J(\frac{\partial\xi}{\partial p})^{-1}=J(\frac{\partial p}{\partial\xi})$となることを用いた．

ここで，$E(\hat{\mathbf{p}})=\mathbf{p}(\Xi)$とするとクラメール・ラオの不等式より

$$Var(\hat{\mathbf{p}})\geq J(\frac{\partial p}{\partial\xi})I(\Xi)^{-1}J(\frac{\partial p}{\partial\xi})' \tag{7}$$

である．一方，$\hat{\mathbf{p}}$は$\mathbf{p}=\mathbf{p}(\Xi)$の不偏推定量と考えることもできるから，不偏推定量の場合のクラメール・ラオの不等式を用いて

$$Var(\hat{\mathbf{p}})\geq I(\mathbf{p})^{-1} \tag{8}$$

でなければならない．式(6)によりどちらで考えてもクラメール・ラオの不等式の下限は一致していることがわかる．

3.4 最尤推定量に基づく信頼区間

未観測の基本式$A_t(\omega)$のi確信度$p_{1_{(x_t)}}$を求めるということは，式(27)で示されているように条件付確率$\mathbf{p}$の要素p_iの部分和を求めることである．本稿では真の母数が未知という立場であるので，最尤推定量$\hat{\Xi}$に基づいて計算された条件付確率$\hat{\mathbf{p}}$の部分和を求めることになる．すなわち，

$$\hat{p}_{1_{(x_t)}}=\sum_{\{X_1,\cdots,X_k\}\setminus\{X_t\}}\hat{p}_{X_1\cdots 1_{(x_t)}\cdots X_k} \tag{9}$$

である．ここで，Nが大きいとき，$\hat{p}_{1_{(x_t)}}$は近似的に平均$p_{1_{(x_t)}}$，分散$Var(p_{1_{(x_t)}})/N$の正規分布に従うとみなせるので，次式が成り立つ．

$$Pr(|\frac{p_{1_{(x_t)}}-\hat{p}_{1_{(x_t)}}}{\sqrt{Var(p_{1_{(x_t)}})/N}}|\leq z_{\alpha/2})\doteq 1-\alpha \tag{10}$$

ただし，$z_{\alpha/2}$は標準正規分布の両側α点である．

本稿の場合，真の母数Ξが未知であると仮定しているため条件付確率の真の値$p_{1_{(x_t)}}$を求めることは不可能である．そこで，分散$Var(p_{1_{(x_t)}})/N$中の$p_{1_{(x_t)}}$を$\hat{p}_{1_{(x_t)}}$に置き換えることにより，式(10)の近似的な信頼区間を求める．この置き換えを行えば$Var(p_{1_{(x_t)}})$は$p_{1_{(x_t)}}$を含まなくなり，式(10)を近似的に$p_{1_{(x_t)}}$に関して解くことにより次式が得られる．

$$\begin{aligned} \hat{p}_{1_{(x_t)}}-\frac{z_{\alpha/2}}{\sqrt{Var(p_{1_{(x_t)}})/N}}<p_{1_{(x_t)}}< \\ \hat{p}_{1_{(x_t)}}+\frac{z_{\alpha/2}}{\sqrt{Var(p_{1_{(x_t)}})/N}} \end{aligned} \tag{11}$$

式(11)が最尤推定量$\hat{\Xi}$に基づく$p_{1_{(x_t)}}$の信頼係数$1-\alpha$の近似的な信頼区間である．

ところでサンプル数Nが大きいとき，最尤推定量Ξの漸近有効性により式(1)の$Var(\hat{\Xi})$は等号を達成することに注意すると，Ξと式(26)(または，(16))の関係にある条件付確率$\hat{\mathbf{p}}$も式(7)の等式を達成するので，分散$Var(\hat{p}_{1_{(x_t)}})$は式(7)の左辺$J(\frac{\partial p}{\partial\xi})I(\Xi)^{-1}J(\frac{\partial p}{\partial\xi})'$を用いることにより求められる．その際，フィッシャー情報行列$I(\Xi)$のΞはその最尤推定量$\hat{\Xi}$で置き換えられる．

3.5 数値例

基本式の数が$3(k=3)$，観測事実の数が$2(s=2)$の場合を考える．今，帰納推論の結果，真の母数Ξの推定値$\hat{\Xi}$，すなわちd結合確信度の値$\hat{\xi}_{x_1x_2x_3}$が各々以下のように与えられたとする．

$$\begin{aligned} &\hat{\xi}_{111}=1.000\times10^{-1},\ \hat{\xi}_{110}=5.000\times10^{-2},\\ &\hat{\xi}_{101}=4.000\times10^{-2},\ \hat{\xi}_{100}=7.000\times10^{-2},\\ &\hat{\xi}_{011}=2.000\times10^{-1},\ \hat{\xi}_{010}=1.000\times10^{-1},\\ &\hat{\xi}_{001}=1.600\times10^{-1},\ \hat{\xi}_{000}=2.800\times10^{-1} \end{aligned} \tag{12}$$

ここで，Ξの7×7のフィッシャー情報行列$I(\Xi)$は以下で与えられる．

$$I(\Xi)=I_1(\Xi)+I_2(\Xi) \tag{13}$$

ここで，$I_1(\Xi)$は対角要素のみが$\frac{1}{\xi_{x_1x_2x_3}}$(ただし，$x_1x_2x_3$は001～111)で他の要素は0である$7\times7$正方行列であり，$I_2(\Xi)$は全ての要素が$\frac{1}{\xi_{000}}$の$7\times7$正方行列である．$I(\Xi)$の$\Xi$を式(12)の最尤推定量$\hat{\Xi}$に置き換え，逆行列$I(\Xi)^{-1}$を計算すると以下のようになる．

$$10^{-2}\times\begin{bmatrix} 13.4, & -1.60, & -3.20, & -1.12, & -0.64, & -0.80, & -1.60 \\ -1.60, & 9.00, & -2.00, & -0.70, & -0.40, & -0.50, & -1.00 \\ -3.20, & -2.00, & 16.0, & -1.40, & -0.80, & -1.00, & -2.00 \\ -1.12, & -0.70, & -1.40, & 6.51, & -0.28, & -0.35, & -0.70 \\ -0.64, & -0.40, & -0.80, & -0.28, & 3.84, & -0.20, & -0.40 \\ -0.80, & -0.50, & -1.00, & -0.35, & -0.20, & 4.75, & -0.50 \\ -1.60, & -1.00, & -2.00, & -0.70, & -0.40, & -0.50, & 9.00 \end{bmatrix} \tag{14}$$

次に，個体に対する観測事実として基本式$A_1(\omega)$のi確信度$q_1=0.5$, $A_2(\omega)$のi確信度$q_2=0.75$が得られ，未観測の基本式$A_3(\omega)$のi確信度$\hat{p}_{1_{(x_3)}}$を求める場合の演繹推論を考える．

前稿の演繹推論アルゴリズム[3]を用いると，以下の結果が得られる．

$$\begin{aligned} &\hat{p}_{111}\asymp2.713\times10^{-1},\ \hat{p}_{110}\asymp1.356\times10^{-1},\\ &\hat{p}_{101}\asymp3.384\times10^{-2},\ \hat{p}_{100}\asymp5.923\times10^{-2},\\ &\hat{p}_{011}\asymp2.287\times10^{-1},\ \hat{p}_{010}\asymp1.144\times10^{-1},\\ &\hat{p}_{001}\asymp5.707\times10^{-2},\ \hat{p}_{000}\asymp9.986\times10^{-2} \end{aligned} \tag{15}$$

この場合，例えばp_{001},p_{011},p_{101},p_{111}は式(26)により各々以下のように書くことができる．

$$\begin{aligned} p_{001}&=\frac{(1-\psi_1)(1-\psi_2)\xi_{001}}{f(\Xi,\Psi)}, & p_{011}&=\frac{(1-\psi_1)\psi_2\xi_{011}}{f(\Xi,\Psi)} \\ p_{101}&=\frac{\psi_1(1-\psi_2)\xi_{101}}{f(\Xi,\Psi)}, & p_{111}&=\frac{\psi_1\psi_2\xi_{111}}{f(\Xi,\Psi)} \end{aligned} \tag{16}$$

ここで，

$$\begin{aligned} f(\Xi,\Psi) &= (1-\psi_1)(1-\psi_2)(\xi_{000}+\xi_{001}) \\ &\quad+(1-\psi_1)\psi_2(\xi_{010}+\xi_{011}) \\ &\quad+\psi_1(1-\psi_2)(\xi_{100}+\xi_{101}) \\ &\quad+\psi_1\psi_2(\xi_{110}+\xi_{111}) \end{aligned}$$

式(16)を$\xi_{x_1x_2x_3}$の各パターンで偏微分することにより，式(7)のヤコビ行列$J(\frac{\partial p}{\partial\xi})$を求めることができる．ここで，式(25)より$q_1$,$q_2$が式(16)の$p_{x_1x_2x_3}$の周辺和であることに注意すると，

$$\begin{aligned} q_1&=p_{100}+p_{101}+p_{110}+p_{111} \\ q_2&=p_{010}+p_{110}+p_{011}+p_{111} \end{aligned} \tag{17}$$

とした連立方程式を解くことによってψ_1，ψ_2を一意に定めることができる．すなわち，式(16)のψ_1,ψ_2の値はq_1,q_2

の値と1対1に対応している．この例の場合，$q_1 = 0.5$, $q_2 = 0.75$と式(12)の値$\hat{\Xi}$を用いると，$\psi_1 = 7.0346 \times 10^{-1}$, $\psi_2 = 7.6226 \times 10^{-1}$と定められる．これらの値を用いてヤコビ行列$J(\frac{\partial p}{\partial \xi})$を計算すると以下のようになる．

$$10^{-2} \times \begin{bmatrix} 33.6, & -6.53, & -6.53, & -4.83, & -4.83, & -15.5, & -15.5 \\ -4.08, & 101, & -13.1, & -9.68, & -9.68, & -31.0, & -31.0 \\ -8.18, & -26.2, & 88.2, & -19.4, & -19.4, & -62.0, & -62.0 \\ -2.11, & -6.77, & -6.77, & 79.6, & -5.01, & -16.1, & -16.1 \\ -1.21, & -3.87, & -3.87, & -2.86, & 81.7, & -9.18, & -9.18 \\ -4.84, & -15.5, & -15.5, & -11.5, & -11.5, & 235, & -36.8 \\ -9.68, & -31.0, & -31.0, & -23.0, & -23.0, & -73.6, & 198 \end{bmatrix} \quad (18)$$

式(14)のフィッシャー情報行列の逆行列$I(\Xi)^{-1}$と式(18)のヤコビ行列を用いて$J(\frac{\partial p}{\partial \xi})I(\Xi)^{-1}J(\frac{\partial p}{\partial \xi})'$を計算すると以下のようになる．

$$10^{-3} \times \begin{bmatrix} 23.2, & 0.56, & 1.11, & 1.29, & 0.74, & -11.5, & -23.0 \\ 0.56, & 122, & -18.4, & -2.74, & -1.57, & -35.2, & -70.5 \\ 1.11, & -18.4, & 225, & -5.48, & -3.13, & -70.5, & -141 \\ 1.29, & -2.74, & -5.48, & 49.7, & -0.21, & -15.9, & -31.7 \\ 0.739, & 1.56, & -3.13, & -0.21, & 28.5, & -9.06, & -18.1 \\ -11.5, & -35.2, & -70.5, & -15.9, & -9.06, & 297, & -141 \\ -23.0, & -70.5, & -141, & -31.7, & -18.1, & -141, & 453 \end{bmatrix} \quad (19)$$

ここで，分散$Var(\hat{p}_{1_{(x_3)}})$は式(7)の等号を満たし，式(19)の行列を用いて

$$\begin{aligned} Var(p_{1_{(x_3)}}) &= Var(p_{001}) + Var(p_{011}) + Var(p_{101}) \\ &\quad + Var(p_{111}) + 2\{Cov(p_{001}, p_{011}) \\ &\quad + Cov(p_{001}, p_{101}) + Cov(p_{001}, p_{111}) \\ &\quad + Cov(p_{011}, p_{101}) + Cov(p_{011}, p_{111}) \\ &\quad + Cov(p_{011}, p_{111})\} = 3.633 \times 10^{-1} \end{aligned} \quad (20)$$

と求められる．

さらに，式(15)の演繹推論アルゴリズムの実行結果より

$$\hat{p}_{1_{(x_3)}} = \hat{p}_{001} + \hat{p}_{101} + \hat{p}_{011} + \hat{p}_{111} = 5.909 \times 10^{-1} \quad (21)$$

が得られるので，$(p_{++1} - \hat{p}_{1_{(x_3)}})/\sqrt{Var(p_{1_{(x_3)}})/N}$は標準正規分布に従うとみなしてよい．今，$\alpha = 0.05$とすると，正規分布表より5パーセント点は$1.96 (= z_{\alpha/2})$であるので，式(10)にこれらの値を代入すると，

$$|\frac{p_{++1} - 0.5909}{\sqrt{0.8955/500}}| \leq 1.96 \quad (22)$$

となる．これをp_{++1}について解くと，

$$5.080 \times 10^{-1} \leq \hat{p}_{++1} \leq 6.739 \times 10^{-1} \quad (23)$$

を得る．

4 まとめ

本稿では，多項分布の母数の推定問題を帰納推論とし，ある確率変数に対する広義の条件付確率の計算問題を演繹推論ととらえることにより，不確実な知識を用いた推論を帰納・演繹推論の二段階に分けて論じた．特に，帰納・演繹の両推論をドメイン全体に対する不確実性と個体に対する不確実性という観点から整理すると，帰納推論ではドメイン全体に対する不確実性しか扱っていないのに対し，演繹推論では両方の不確実性を切り分けて扱うことにより条件付確率の計算に落とし込んでいる点が重要である．この条件付確率の計算は，ベイズ統計学的な視点からとらえると事後確率の計算とみなすこともできる．

さらに，確信度に幅を持たせることにより演繹推論結果の信頼性を保証する推論法を構築した．本稿で扱った問題は，帰納推論結果の最尤推定量$\hat{\Xi}$を用いて演繹推論を行い，条件付確率$\hat{\mathbf{p}}$を計算するという二重構造になっている．すなわち，演繹推論時には推定誤差が生じる余地はなく，本質的に推定誤差を含んでいるのは$\hat{\Xi}$のみであり，$\hat{\mathbf{p}}$は$\hat{\Xi}$の推定誤差を引き継ぐ形式のモデルになっている．そこで，最尤推定量の漸近有効性と式(1)の多変数の場合のクラメール・ラオの不等式により$\hat{\Xi}$の分散が近似的に最小になるという性質に着目し，式(26)で示されるΞから$\mathbf{p}$へのパラメータ変換によって生じるフィッシャー情報行列の変換を用いて，漸近的かつ近似的に区間の幅を最小にする$p_{1_{(x_3)}}$の確信度区間を求めた．

謝辞

本研究の一部は文部省科学研究費(日本学術振興会特別研究員奨励金)の補助による．

参考文献

[1] Ireland, C.T., and Kullback, S. : Contingency tables with given marginals, *Biometrika* , Vol. 55, 1, pp. 179–188 (1968).

[2] Pearl, J. : *Probabilistic Reasoning in Intelligent Systems* , Morgan Kaufmann (1988).

[3] 鈴木,松嶋,平澤 : 不確実な知識を用いた推論のモデル化と推論法について, 情報処理学会論文誌, Vol. 41, No.1, pp. 1–11 (2000).

[4] 竹村彰通, : 現代数理統計学, 創文社 (1991).

[5] 竹内啓, : 数理統計学, 東洋経済 (1963).

A 付録–前稿の演繹推論の定式化–

ここで，前稿における演繹推論の定式化を以下の式(24)〜(27)に示す§．まず，$X_1, \cdots, X_k, y_{r_1}, \cdots, y_{r_s}$の結合確率分布は式(24)で表現できる．

$$\begin{aligned} &P(X_1, \cdots, X_k, y_{r_1}, \cdots, y_{r_s} \mid \Xi, \Psi) \\ &= P(X_1, \cdots, X_k \mid \Xi) \prod_{i=1}^{s} P(y_i \mid X_1, \cdots, X_k, \Xi, \Psi) \\ &= P(X_1, \cdots, X_k \mid \Xi) \prod_{i=1}^{s} P(y_i \mid X_i, \psi_i) \end{aligned} \quad (24)$$

次に，観測事実であるs個の基本式$A_{r_m}(\omega)$についてのi確信度$q_{1_{(x_{r_m})}}$は，真理値の観測結果y_{r_m}と既知のパラメータΞと未知のパラメータψ_{r_m}が与えられたもとでの真理値x_{r_m}の条件付確率とみなすことができ，式(25)で表現できる．

$$\begin{aligned} q_{1_{(x_{r_m})}} &= \sum_{\substack{\{X_1, \cdots, X_k\} \setminus \{X_{r_m}\} \\ X_{r_m} = 1, \cdots, X_k}} P(X_1, X_2, \cdots, \\ &\quad \mid y_{r_1}, \cdots, y_{r_s}, \Xi, \Psi) \\ &= P(X_{r_m} = 1 \mid y_{r_m}, \psi_{r_m}) \end{aligned} \quad (25)$$

さらに，条件付確率(i結合確信度)は式(26)のように表現できる．

$$\begin{aligned} p_{x_1 \cdots x_k} &= P(X_1 = x_1, \cdots, X_k = x_k \mid y_{r_1}, \cdots, y_{r_s}, \Xi, \Psi) \\ &= \frac{P(X_1, \cdots, X_k, y_{r_1}, \cdots, y_{r_s} \mid \Xi, \Psi)}{P(y_{r_1}, \cdots, y_{r_s} \mid \Xi, \Psi)} \\ &= \frac{P(X_1, \cdots, X_k \mid \Xi) \prod_{i=1}^{s} P(y_{r_i} \mid X_{r_i}, \psi_i)}{\sum_{X_1, \cdots, X_k} \{P(X_1, \cdots, X_k \mid \Xi) \prod_{i=1}^{s} P(y_{r_i} \mid X_{r_i}, \psi_i)\}} \end{aligned} \quad (26)$$

最後に，演繹推論結果である基本式$A_t(\omega)$のi確信度$p_{1_{(x_t)}}$は式(27)で表現できる．

$$\begin{aligned} p_{1_{(x_t)}} &= \sum_{\{X_1, \cdots, X_k\} \setminus \{X_t\}} P(X_1, \cdots, X_t = 1, \cdots, X_k \mid \\ &\quad y_{r_1}, \cdots, y_{r_s}, \Xi, \Psi) \\ &= \sum_{\{X_1, \cdots, X_k\} \setminus \{X_t\}} p_{X_1 \cdots 1_{(x_t)} \cdots X_k} \end{aligned} \quad (27)$$

§ 詳細は前稿[3]を参照されたい．

The 23rd Symposium on Information Theory
and Its Applications (SITA2000)
Aso, Kumamoto, Japan, Oct. 10 - 13, 2000

ユニバーサル符号における固定長符号の誤り率と可変長符号のオーバーフロー確率について

On the error probability of FF codes and the overflow probability of FV codes in universal source coding

野村 亮* 松嶋 敏泰* 平澤 茂一*
Ryo NOMURA Toshiyasu MATSUSHIMA Shigeichi HIRASAWA

Abstract— In the lossless source coding, it was shown that the relationship of the error probability of FF codes and the overflow probability of FV codes for general sources, when the source distribution is known. In this paper, first we show the relationship of the error probability of FF codes and the overflow probability of FV codes in universal source coding. Then we show the infimum rate of the code, which minimizes the error probability with respect to the Bayes criteria, under the condition that the error probability goes to 0.

Keywords— universal source coding,FF Codes,FV Codes, asymptotic normality

1 はじめに

歪のない情報源符号化において，従来多くの符号は平均符号長を基準として研究されてきた．それに対して，Merhavは近年オーバーフロー確率を定義し，それに関する解析を行った[1]．Merhavの定義したオーバーフロー確率とは，一つの系列を符号化した際にその1シンボルあたりの符号長がある定数を超える確率を表わしており，可変長符号における一つの評価基準である．一方，固定長符号における評価基準として誤り確率の評価があげられる．近年この二つの関係が論じられており，情報源が既知の場合，任意の一般情報源に対して固定長符号における誤り確率の最小達成可能指数レートと可変長符号におけるオーバーフロー確率の最小達成可能指数レートが等しいことが示されている[2]．

本研究では，情報源の確率構造が未知の場合を考える．まず，固定長符号における誤り確率の達成可能な最小レートと可変長符号におけるオーバーフロー確率の達成可能な最小レートが等しいことを示す．次に，ベイズ基準のもとで，固定長符号における誤り確率を最小化する符号化を提案する．さらに，ベイズ基準のもとで，固定長符号の誤り確率を最小化することと可変長符号においてオーバーフロー確率を最小化することが等価であることを示す．最後に，ベイズ基準のもとで誤り確率を最小化する固定長符号の達成可能最小レートを漸近的に評価する．

2 準備

本研究では，定常エルゴード情報源を対象とし，情報源アルファベットを$A=\{i:0\leq i\leq d-1\}$とする．符号cを用いてx^nを符号化したときの符号長を$l(c(x^n))$とすると，無歪み符号cは以下を満たす必要がある[3]

$$\sum_{x^n\in X^n} 2^{-l(c(x^n))}\leq 1. \tag{1}$$

なお，本研究では符号長を理想符号長$-\log q(x^n)$で考える．

3 固定長符号における誤り率と可変長符号におけるオーバーフロー確率の関係

ある可変長符号φ^Vに対して，情報源系列x^nに対する符号長を$l(\varphi^V(x^n))$としたとき，以下の確率をオーバーフロー確率と呼ぶ[4]．

$$\eta_n=\Pr\left\{l(\varphi^V(x^n))>\epsilon_n\right\}. \tag{2}$$

本研究ではオーバーフロー確率を評価する際，次の量を考える．

定義 3.1

$$\lim_{n\to\infty}\eta_n=0, \tag{3}$$

となるユニバーサル符号が存在するとき，ϵ_nを可変長ユニバーサル達成可能という．ここで，ϵ_nはnに関して単調増加なある関数である． □

さらに，

定義 3.2

$$L^*_{\varphi^V}=\{\inf\epsilon_n|\epsilon_n\text{が可変長ユニバーサル達成可能}\}, \tag{4}$$

をユニバーサル可変長達成可能最小レートと呼ぶ． □

とする．

次に固定長符号における誤り確率を定義する．ある固定長符号$\varphi^F(\cdot)$に対して，情報源系列x^nに対する符号長を$\log M_n$としたとき，以下の確率を固定長符号における誤り確率と呼ぶ．

$$\nu_n=\Pr\left\{x^n\neq\varphi^{F^{-1}}\varphi^F(x^n)\right\} \tag{5}$$

ここで次の量を定義する．

定義 3.3

$$\lim_{n\to\infty}\nu_n=0, \tag{6}$$

かつ，

$$\lim_{n\to\infty}\log M_n-\epsilon_n<0, \tag{7}$$

となるユニバーサル固定長符号が存在するとき，ϵ_nを固定長ユニバーサル達成可能という． □

さらに，

* 早稲田大学理工学部経営システム工学科，〒169-8555 新宿区大久保3-4-1. Department of Industrial and Management Systems Engineering, School of Science and Engineering, WASEDA University, 3-4-1, Ohkubo, Shinjuku-ku, Tokyo. E-mail:ryochan@matsu.mgmt.waseda.ac.jp

定義 3.4

$$R^*_{\varphi^F} = \{\inf \epsilon_n | \epsilon_n \text{が固定長ユニバーサル達成可能}\}, \quad (8)$$

をユニバーサル固定長達成可能最小レートと呼ぶ. □

とする.

注意 3.1 従来の情報源符号化においては, 固定長符号の誤り率をν_nとしたとき

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\nu_n} \geq r, \quad (9)$$

となるもとで

$$\lim_{n\to\infty} \frac{1}{n} \log M_n \leq R, \quad (10)$$

となる符号が存在するときレートRを達成可能と呼んだ[2]. これは誤り率が指数的に0に向かうことを要請したものである. それに対して, 本研究では誤り率が0に向かうという条件のみで指数的に0に向かうことを要請していないことに注意されたい. □

次の定理はユニバーサル固定長達成可能最小レートとユニバーサル可変長達成可能最小レートの関係を示している.

定理 3.1 任意の定常エルゴード情報源に対して

$$L^*_{\varphi^V} = R^*_{\varphi^F}, \quad (11)$$

が成立する. □

(証明)

$L^*_{\varphi^V} \leq R^*_{\varphi^F}$と$L^*_{\varphi^V} \geq R^*_{\varphi^F}$を証明すればよい.

a)$L^*_{\varphi^V} \leq R^*_{\varphi^F}$の証明

ϵ_nが固定長ユニバーサル達成可能であるとする. すなわち

$$\lim_{n\to\infty} \nu_n = r, \quad (12)$$

かつ

$$\lim_{n\to\infty} \log M_n - \epsilon_n < 0, \quad (13)$$

となるFF符号φ^Fが存在する. この符号をφ^F_Cとする.

ここで, 以下のような集合を定義する.

$$T_n = \left\{x^n \in \mathcal{X}^n | x^n = \varphi^{F^{-1}}(\varphi^F_B(x^n))\right\} \quad (14)$$

さらに, 次のようなFV符号器を用意する.

$$\varphi^V_C(x^n) = \begin{cases} 1 \ * \ \varphi^F_C(x^n) & x^n \in T_n \\ 0 \ * \ FV\text{符号器} & otherwise, \end{cases} \quad (15)$$

ここで, $*$は系列の連節[1]を表す.

上記で定義されたFV符号器のオーバーフロー確率を考える. ϵ_nが単調増加であることより, 十分大きいnに対して$x^n \in T_n$であれば

$$\begin{aligned} l(\varphi^V_C(x^n)) &= \log M_n \\ &< \epsilon_n, \end{aligned} \quad (16)$$

である. ゆえに

$$\Pr\{l(\varphi^V_C(x^n)) > \epsilon_n\} \leq \Pr\{x^n \notin T_n\}, \quad (17)$$

であることがわかる. ここで, T_nの定義より

$$\lim_{n\to\infty} \Pr\{x^n \notin T_n\} = 0, \quad (18)$$

であるので結局$L^*_{\varphi^V} \leq R^*_{\varphi^F}$が成立する.

b)$L^*_{\varphi^V} \geq R^*_{\varphi^F}$の証明

ϵ_nが可変長ユニバーサル達成可能であるとする. 次のような集合を定義する.

$$S_n = \left\{x^n \in \mathcal{X}^n | l(\varphi^V(x^n)) \leq \epsilon_n\right\}. \quad (19)$$

すると, ϵ_nが可変長ユニバーサル達成可能であることより, 明らかに

$$\lim_{n\to\infty} \Pr\{x^n \notin S_n\} = 0, \quad (20)$$

である. また$M_n = |S_n|$とし, 次のような固定長符号器を考える.

$$\varphi^F(x^n) = \begin{cases} 1, 2, \cdots, M_n, & x^n \in S_n \\ 1, & otherwise \end{cases} \quad (21)$$

すると明らかに

$$M_n = |S_n| \leq d^{n\epsilon_n}, \quad (22)$$

であるので

$$\lim_{n\to\infty} \log M_n - \epsilon_n < 0, \quad (23)$$

が成立する. また, 誤り率η_nは

$$\eta_n = \Pr\left\{x^n \neq \varphi^{F^{-1}}(\varphi^F(x^n))\right\} = \Pr\{x^n \notin S_n\}, \quad (24)$$

となる. 上式と(20)式より, $L^*_{\varphi^V} \geq R^*_{\varphi^F}$がわかる. □

この定理より固定長ユニバーサル達成可能最小レートは可変長ユニバーサル達成可能最小レートと等しいことがわかる.

4 誤り率最小固定長符号

本節ではベイズ基準の下で固定長誤り率ν_nを最小にする符号を考える. ここで, 情報源系列の出現確率はパラメータ$\theta \in \Theta$に従うパラメトリックなモデルで, θは未知とする. このもとで, 次のような符号化法を考える.

step1 $m(x^n) = \int_{\theta\in\Theta} P(x^n|\theta)p(\theta)d\theta$を計算.

step2 $m(x^n)$の大きいほうからM_n個の系列の集合をB_nとする.

step3 $x^n \in B_n$は長さ$\log M_n$の系列へ符号化.
それ以外は長さ$\log M_n$の系列の一点へ符号化.

定理 4.1 上記の符号化法はベイズ基準のもとで誤り率を最小にする. □

(証明)

損失関数を$\nu_n = \Pr\left\{x^n \neq \varphi^{F^{-1}}(\varphi^F(x^n))\right\}$とすると, ベイズリスクは以下のようになる.

$$\int_{\theta\in\Theta} \Pr\left\{x^n \neq \varphi^{F^{-1}}(\varphi^F(x^n))\right\} P(\theta)d\theta. \quad (25)$$

ここで, ある固定長符号$\varphi^F(\cdot)$において誤って復号される系列の集合をC^φ_nとする. すなわち

$$C^\varphi_n = \left\{x^n \in \mathcal{X}^n | x^n \neq \varphi^{F^{-1}}(\varphi^F(x^n))\right\}, \quad (26)$$

[1] 例えば,$01 * 010 = 01010$である.

である．するとベイズリスク最小の符号化は

$$\begin{aligned}
&\min_{\varphi}\int_{\theta\in\Theta}\Pr\left\{x^n\neq\varphi^{F^{-1}}(\varphi^F(x^n))\right\}P(\theta)d\theta\\
&=\min_{\varphi}\int_{\theta\in\Theta}\sum_{x^n\in C_n^{\varphi}}P(x^n|\theta)P(\theta)d\theta\\
&=\min_{\varphi}\sum_{x^n\in C_n^{\varphi}}\int_{\theta\in\Theta}P(x^n|\theta)P(\theta)d\theta\\
&=\min_{\varphi}\sum_{x^n\in C_n^{\varphi}}m(x^n),
\end{aligned}\tag{27}$$

と書ける．□

さらにベイズ基準のもとでオーバーフロー確率最小の可変長符号も同様に定義できる．すなわち，あるパラメータθのもとでオーバーフロー確率を次のように定義する．

$$\eta_n=\Pr\left\{l(\varphi^V(x^n))>\epsilon_n\right\}.\tag{28}$$

この量に対して，同様のパラメトリックな情報源を考えたとき以下の量を最小化する符号化である．

$$\int_{\theta\in\Theta}\Pr\left\{l(\varphi^V(x^n))>\epsilon_n\right\}P(\theta)d\theta.\tag{29}$$

ここで

$$C_n^{\varphi}=\left\{x^n\in\mathcal{X}^n|x^n\neq\varphi^{F^{-1}}(\varphi^F(x^n))\right\},\tag{30}$$

とおくと，次の定理が成立する．

定理 4.2 $C_n^{\varphi}=\{x^n|-\log m(x^n)<\epsilon_n\}$としたとき，次が成り立つ．

$$\begin{aligned}
&\min_{L(\cdot)}\int_{\theta\in\Theta}\Pr\left\{l(\varphi^V(x^n))>\epsilon_n\right\}P(\theta)d\theta\\
&=\min_{\varphi}\int_{\theta\in\Theta}\Pr\left\{x^n\neq\varphi^{F^{-1}}(\varphi^F(x^n))\right\}P(\theta)d\theta
\end{aligned}\tag{31}$$

(証明)
(29)の最小化は

$$\begin{aligned}
&\min_{L(\cdot)}\int_{\theta\in\Theta}\Pr\left\{L^{\varphi^V}(x^n)>\epsilon_n\right\}P(\theta)d\theta\\
&=\min_{q(\cdot)}\int_{\theta\in\Theta}\Pr\left\{-\log q(x^n)>\epsilon_n\right\}P(\theta)d\theta\\
&=\min_{q(\cdot)}\int_{\theta\in\Theta}\sum_{-\log q(x^n)>\epsilon_n}p(x^n|\theta)P(\theta)d\theta\\
&=\min_{q(\cdot)}\int_{\theta\in\Theta}\sum_{q(x^n)<e^{-\epsilon_n}}p(x^n|\theta)P(\theta)d\theta\\
&=\min_{q(\cdot)}\sum_{q(x^n)<e^{-\epsilon_n}}m(x^n)\\
&=\min_{D_n}\sum_{x^n\in D_n}m(x^n),
\end{aligned}\tag{32}$$

となる．ただし，$D_n=\{x^n|q(x^n)<e^{-\epsilon_n}\}$とおいた．ここで，(27)式における$C_n\equiv D_n$とおくと定理が証明される．□

上記の定理より，ベイズ基準のもとで固定長符号の誤り率を最小化することと可変長符号のオーバーフロー確率を最小化することは等価であることがわかる．このことより，固定長符号の誤り率最小符号に関する性質は可変長符号におけるオーバーフロー確率最小符号に適用できることがわかる．

5 固定長符号誤り率の漸近評価

本節では，前節で定義したベイズ基準のもとでの固定長符号の誤り率を最小にする符号（以下ベイズ誤り率最小符号と呼ぶ）の誤り率を評価する．ベイズ誤り率最小符号の符号化関数を$\varphi_B^F(\cdot)$とし以下を定義する．

定義 5.1

$$\lim_{n\to\infty}\Pr\left\{x^n=\varphi_B^{F^{-1}}(\varphi_B^F(x^n))\right\}=0,\tag{33}$$

かつ，

$$\lim_{n\to\infty}\frac{1}{J_n}\log M_n-\frac{\epsilon_n}{J_n}<0,\tag{34}$$

となるとき，ベイズ誤り率最小符号で達成可能であるという．ここで，J_nは単調非減少なnの関数である．□

注意 5.1 前節までの達成可能の定義と異なっていることに注意されたい．□

さらに，次を定義する．

定義 5.2

$$R_B^*=\inf\{\epsilon_n|\epsilon_n\text{がベイズ誤り率最小符号で達成可能}\},\tag{35}$$

をベイズ誤り率最小符号達成可能最小レートと呼ぶ．□

ここで，次を仮定する．

仮定 5.1 $-\log m(x^n)$が漸近正規性を満たす．すなわち

$$\frac{-\log m(x^n)-E[-\log m(X^n)]}{\sqrt{nV_B^2}}\sim N(0,1),\tag{36}$$

が成立する．ここで，V_B^2は次式で定義されるものでベイズ符号の符号長の分散と呼ばれる *[4][5]*.

$$V_B^2=\lim_{n\to\infty}\frac{1}{n}E\left[\{-\log m(x^n)-E[-\log m(X^n)]\}^2\right].$$

□

上記の仮定のもとで次の定理が成り立つ．

定理 5.1 仮定*5.1*のもとで，任意の$J_n>O(\sqrt{nV_B^2})$に対して

$$R_B^*=nH(X)+O(\sqrt{nV_B^2}),\tag{37}$$

が成立する．

(証明) 付録参照．□

上記の定理より，ベイズ誤り率最小符号を用いたとき誤り率が漸近的に0になるために必要なレートにはベイズ符号の符号長の分散が重要であることがわかる．

6 まとめ

本研究では，固定長符号の誤り率が漸近的に0になるために必要な最小レートが可変長符号のオーバーフロー確率が漸近的に0になるために必要な最小レートと等しいことを情報源が未知の場合に関して示した．さらに，ベイズ基準のもとで，固定長符号の誤り率を最小化することはオーバーフロー確率を最小化することと等価であることを示した．このことにより，ベイズ基準のもとでの固定長符号誤り率最小符号に関する定理はオーバーフロー確率最小符号へと適用できることがわかる．最後にベイズ基準のもとで固定長符号の誤り率最小符号の誤り率が0になるために必要な最小レートをベイズ符号の符号長の分散を用いて示した．

謝辞

本研究を行うにあたり，御討論，ご助言頂いた平澤・松嶋両研究室内各氏に深く感謝いたします．なお，本研究の一部は文部省科学研究費基盤(C)(No.12650400)，早稲田大学特定課題研究助成費(99A-551)の援助による．

参考文献

[1] N.Merhav, "Universal Coding with Minimum Probability of Codeword Length Overflow," *IEEE Trans. Inf. Theory*,37(3):556-563,1991.

[2] O.Uchida, "An Information-Theoretic Study on Variable-Length Source Coding with Unequal Cost, " 電気通信大学博士論文,2000.

[3] T.M.Cover and J.A.Thomas, "Elements of information theory," Wiley,1991.

[4] R.Nomura, T.Matsushima and S.Hirasawa, "On the variance and the probability of length overflow of lossless codes," In *Proc.Int.Symp.on.Inf.Theory*, page 44,2000.

[5] 野村亮，松嶋敏泰，平澤茂一，"符号長の分散とオーバーフロー確率について，" 第22回情報理論とその応用シンポジウム予稿集,pp.355-358,1999.

付録 定理5.1の証明

まず，二つの集合を定義する．

$$T_n = \left\{ x^n \,\middle|\, x^n = \varphi_B^{F^{-1}}(\varphi_B^F(x^n)) \right\}. \tag{38}$$

さらに，$J_n > O(\sqrt{nV_B^2})$ として

$$K_n = \{x^n \,||-\log m(x^n) + E\left[-\log m(x^n)\right]| \le J_n\}, \tag{39}$$

とする．すると仮定5.1より，$-\log m(x^n)$ は漸近正規性を満たしているので，次がわかる[4][5].

$$\lim_{n\to\infty} \Pr\{K_n\} = 1, \tag{40}$$

かつ，任意の $J_n > O(\sqrt{nV_B^2})$ に対して

$$\lim_{n\to\infty} \Pr\{x^n \in K_n\} = 1, \tag{41}$$

である．また，K_n の定義と $-\log m(x^n)$ の漸近正規性より，十分大きい n のもとで $\forall x^n \in K_n$ に対して

$$\begin{aligned} & \frac{-\log m(x^n) + O(\sqrt{nV_B^2})}{J_n} - \frac{E[-\log m(x^n)]}{J_n} \\ = & \frac{-\log m(x^n) + O(\sqrt{nV_B^2})}{J_n} - \frac{nH(X)}{J_n} \\ < & \gamma, \end{aligned} \tag{42}$$

が成り立つ．ここで最初の等式はベイズ符号の平均符号長から成立する．また，$\forall \gamma > 0$ である．

このもとでまず，$|T_n| = nH(X) + O(\sqrt{nV_B^2})$ がベイズ誤り率最小符号で達成可能であることを示す．

K_n の定義より $\forall x^n \in K_n$ に対しては

$$\begin{aligned} 1 \ge & \sum_{x^n \in K_n} m(x^n) \\ \ge & |K_n| \exp\left\{ -nH(X) - O(\sqrt{nV_B^2}) \right\}, \end{aligned} \tag{43}$$

であるので

$$|K_n| \le \exp\left\{ nH(X) + O(\sqrt{nV_B^2}) \right\}, \tag{44}$$

となる．結局

$$\log |K_n| \le nH(X) + O(\sqrt{nV_B^2}), \tag{45}$$

となる．すると符号の構成法より明らかに

$$K_n \subseteq T_n, \tag{46}$$

であるので，式(40)より $nH(X) + O(\sqrt{nV_B^2})$ はベイズ誤り率最小符号で達成可能である．

次に，任意の $|T_n| < nH(X) + O(\sqrt{nV_B^2})$ がベイズ誤り率最小符号で達成可能でないことを示す．

(42)式より，$\forall x^n \in K_n$ に対して

$$\lim_{n\to\infty} \left| \frac{-\log m(x^n)}{J_n} - \frac{nH(X) + O(\sqrt{nV_B^2})}{J_n} \right| = 0, \tag{47}$$

である．これより

$$\begin{aligned} & \lim_{n\to\infty} \frac{1}{m(x^n)\exp\{J_n\}} \\ = & \lim_{n\to\infty} \exp\left\{ \frac{nH(X) + O(\sqrt{nV_B^2})}{J_n} \right\} \\ = & \lim_{n\to\infty} |K_n| \exp\{-J_n\}, \end{aligned} \tag{48}$$

である．ゆえに

$$\lim_{n\to\infty} \left| \frac{1}{J_n} \log |K_n| - \frac{nH(X) + O(\sqrt{nV_B^2})}{J_n} \right| = 0, \tag{49}$$

となる．一方

$$|T_n| < nH(X) + O(\sqrt{nV_B^2}), \tag{50}$$

であるので明らかに $K_n \not\subseteq T_n$, である．ここで

$$\begin{aligned} & \Pr\left\{ x^n = \varphi_B^{F^{-1}}(\varphi_B^F(x^n)) \right\} \\ = & \Pr\{x^n \notin T_n\} \\ > & \Pr\{x^n \in K_n \setminus T_n\}, \end{aligned} \tag{51}$$

であり，また

$$\lim_{n\to\infty} \Pr\{x^n \in K_n \setminus T_n\} > 0, \tag{52}$$

である．ゆえに T_n はベイズ誤り率最小符号で達成可能でない． □

The 23rd Symposium on Information Theory and Its Applications (SITA2000)
Aso, Kumamoto, Japan, Oct. 10 - 13, 2000

An Iterative Calculation Algorithm for Posterior Probability *

Toshiyasu MATSUSHIMA † Tomoko K. MATSUSHIMA ‡ Shigeichi HIRASAWA †

Abstract— An iterative procedure that calculates an exact posterior marginal probability is proposed in this paper. The basic concept of the proposed algorithm is provided by a differential geometrical interpretation on the calculation of posterior probability. It is shown that an exact posterior probability can be calculated by the e-projections of an initial probability on a submanifold satisfying a given condition. The applicable probability model class of the proposed algorithm is wider than that of the Belief Propagation (BP) algorithm. The probability model class for which the algorithm guarantees exact answers is also wider than that of the BP algorithm.

Keywords— Posterior probability, Differential geometry, Bayesian network, Belief propagation, Turbo decoding

1 Introduction

Let $\mathbf{u} = (u_1, \cdots, u_K)$, $u_k \in A$ be a information sequence and $\mathbf{y}$ a received sequence. The information sequence $\mathbf{u}$ is supposed to be encoded systematically a codeword$(\mathbf{u}, \mathbf{x})$, where $\mathbf{u}$ is the systematic part and $\mathbf{x}$ is the parity part.

The decoding problem is defined as the estimation of the information sequence $\hat{\mathbf{u}}$ from the received sequence $\mathbf{y}$. For minimizing symbol error probability, we select the information symbol $\hat{u}_k$ that maximize the posterior probability $P(u_k|\mathbf{y})$ given the received sequence $\mathbf{y}$. Thus the calculation of the posterior marginal probability $P(U_k|\mathbf{y})$ is considered as a main problem in the optimal symbol-by-symbol decoding. The turbo decoding algorithm is an iterative algorithm calculating the posterior probability approximately.

In several research fields, the probability structure on random variables is represented by using graph. Bayesian Network is a typical graphical representation in the field of AI[6][10]. The conditional distribution of a event is represented on the directed arc from its parents node to the node of the event in BN.

In the field of knowledge representation using BN, the fundamental probabilistic inference problem is to compute the update beliefs, i.e., the posterior probability of an event given some evidences. The belief of each node or event is updated by propagation of the messages π and λ. This update algorithm is called Belief Propagation (BP).

The BP algorithm can not be applied to all BN. The algorithm is only applied to polytrees which have no loop in BN. The BP algorithm guarantees exact results for polytrees [10].

There are several studies on the graphical representation for the probability structure of codes. Tanner graph[4][5] is an early research of graphical representation of code. Recently, representing some codes by using BN, the relation between turbo decoding and Belief Propagation is investigated[4][5][7].

Since the BN representing compound codes, which include the original turbo codes, have loops, the procedure of BP can not be used directly for decoding turbo codes. In the BN that has loops, the BN is broken into tractable subnetworks that have no loops, and the BP algorithm is reciprocally applied to each subnetwork. The compound BP algorithm such as the turbo decoding do not calculate necessarily correct answers.

In this paper, we shall propose an iterative algorithm that calculate marginal posterior probabilities from the view point of differential geometry. There are some researchies[1][9] that applies differential geometry to statistics. The calculation of posterior probability under a certain condition is interpreted as the e-projection of a prior probability on the manifold satisfying the given condition. If multi-conditions are given, the posteior probability is given by repeating e-projection iteratively to each condition.

The applicable probability model class of the proposed algorithm is wider than that of the BP algorithm. The probability model class for which the proposed algorithm caluculates exact marginal posterior probability is also wider than that of the BP algorithm.

The proposed algorithm can be used many application fields such as turbo decoding. Although the primitive BP algorithm can not be applied to the compound code decoding directly, the proposed algorithm can be directly applied to the probability model of the compound codes. The algorithm is not necessary to break the model, i.e., the BN, into tractable submodels, i.e., subnetworks.

2 Probability model class and problem

We shall investigate a probability model class in this paper. The model class is called log linear models[12][13]. The joint probability of a log linear model is represented by the following function.

$$\begin{aligned}\log P(x_1, \ldots, x_n) &= \mu + q(x_1) + \cdots + q(x_n) \\ &\quad + \cdots + q(x_i, x_j) + \cdots + q(x_i, x_j, x_k) \\ &\quad + \cdots + q(x_1, \ldots, x_n), \end{aligned} \tag{1}$$

where

$$\sum_{x_i} q(x_1) = 0,$$

* This research was supported in part by the Ministry of Education under Grant-Aids (c) No.12650400 for Scientific Research and Waseda University under Grant 99A-551 for Special Research Projects.

† School of Science and Engineering, Waseda University, 3-4-1 Ohkubo, Shinjuku-ku, Tokyo, 169-8555 JAPAN, E-mail: toshi@matsu.mgmt.waseda.ac.jp

‡ Dept. of Information Engineering, Polytechnic University, 4-1-1 Hashimotodai, Sagamihara, 229-1196 JAPAN

$$\sum_{x_i} q(x_i, x_j) = \sum_{x_j} q(x_i, x_j) = 0,$$

$$\sum_{x_i} q(x_i, x_j, x_k) = \sum_{x_j} q(x_i, x_j, x_k) = \sum_{x_i} q(x_i, x_j, x_k) = 0 \quad (2)$$

This probability model class includes the model class represented by BN.

We shall introduce several notations. A random variable X_i is represented by the index i for short. A set of the indices of some random variables is denoted by I. Then, a term $q(X_i, X_j)$ in formula (1) is represented by $q(i, j)$ or $q(I)$. A probability distribution $P(X_i, X_j)$ is also abbreviated to $P(i, j)$ or $P(I)$.

Let the set of the indices of all random variables in a given probability model be denoted by I^n. The problem that we shall investigate is defined as follows.

Problem 2.1 *Calculate the marginal posterior probabilities $P(i|C)$ $(i \in I^n)$ under the condition C that restrics the marginal probabilities $P^*(i)$ $(i \in I^k \subset I^n)$.*

If the given marginal probability is a point mass in $X_i = x_i$ $(i \in I^k)$ as $P^*(X_i = x_i) = 1$ $(i \in I^k)$ then $P(i|C)$ is the ordinary posterior probability $P(X_i|X_i(1) = x_i(1), \ldots, X_i(k) = x_i(k))$.

Scince the calculation of the posterior marginal probability $P(U_i|\mathbf{Y} = \mathbf{y})$ is a main problem in the optimal symbol-by-symbol decoding, Problem 2.1 includes an important problem in decording. Problem 2.1 is also an important problem in the AI field, since Problem 2.1 is exactly the fundamental probabilistic inference problem[11] that is to compute the update beliefs, i.e., the marginal posterior probability of an event given some evidences.

3 The proposed algorithms

We shall proposed an algorithm for calculating Problem 2.1. The algorithm consists of two procedures. The pre-procedure calculates maximal terms, which are used as the units to calculate posterior probability in the main-procedure. Although the main-procedure does not use a graphical model such as BN explicitly, the main term set is used for representing a given probability model.

First, we shall explain the pre-procedure. A term $q(I)$ of a log linear model is represented by a set of indices I. The size of a term $q(I)$ is defined as the number of the elements in the index set I. Let T be the set of the index sets of all term in a given log linear model. Let $V \subset T$ be the set of the index sets of all maximal term in a given log linear model. The pre-procedure calculates the maximal term set V from T.

[Pre-procedure]
begin
$T_R := T$;
$V := \phi$;
while $T_R \neq \phi$ do
  begin
  Pick up the term $q(I)$ whose size is largest in T_R;
  $T_R := T_R - \{I\}$;
  if $\forall_{J \in V}\, I \not\subset J$ then
  $V := V \cup \{I\}$;
  end
end

Secondly, we shall show the main-procedure that calculates the marginal posterior probabilities in Problem 2.1.

[Main-procedure]
begin
while $\exists_{I \in V}$ new $P(I) \neq$ previous $P(I)$ do
  begin
  Pick up an index i from I^k;
  $T_0 := \{I | i \in I, I \in V\}$;
  $S_0 := T_0$;
  $I_i := i$;
  Put T_0 in the list L_i
  The set $C := \phi$;
  while $L \neq \phi$ do
    begin
    Pick up T_i that is the top element of L_i
    $C := C \cup T_i$;
    for all $I_j \in T_i$ do
      begin
      Calculate $P(I'')$ from $P(I_j)$
      where $I'' = I_i \cap I_j$;
      $P(I_j) :=$ the e-projection of $P(I_j)$ on
      the manifold satisfying $P(I'')$;
      $T_j := \{I | I_j \cap I \neq \phi \wedge I \notin C, I \in V - S_i\}$;
      $S_j := I_j$;
      Put T_j on the tail of the list L_i;
      end
    end
  end
end

We can use the various order for picking up an index i from I^k at line 4. An example is numerical order such as $i(1), i(2), \ldots, i(k), i(1), \cdots i(j) < i(j+1))$. The e-projection of $P(I_j)$ to the manifold satisfying $P(I'')$ is easily calculated by Iterative Scaling Procedure(ISP)[11]. We say that a collision occurs when an I is included in C at line 20.

4 Projection and posterior probability calculation

For the proof of the justice of the proposed algorithm, we shall investigate the calculation of the posterior probability from the viewpoint of differential geometry.

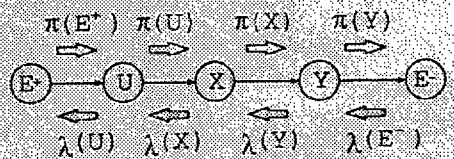

Figure 1: A probability model represented by BN

Lemma 4.1 *Let Z be random variables and t be a measurable mapping of Z into $t(Z)$. The t-image of a probability measure q on Z is denoted by q^t. Given a measure q^{t*} on $t(Z)$, define $S_q = \{q|q^t = q^{t*}\}$ and a probability measure p on Z. The posterior probability q^* given the information of $q^t = q^{t*}$ is given by*

$$q^* = \arg\min_{q \in S_q} D^{(t-1)}(q\|p). \tag{3}$$

Especially, if q^{t*} is the point mass in $t(Z) = t$, q^* is the ordinary conditional probability $P(z|t(Z) = t)$ given t.

Assume $Z = (\mathbf{U}, \mathbf{X}, \mathbf{Y})$ and $t(Z) = \mathbf{Y}$. The posterior probability $P(\mathbf{U}|\mathbf{Y} = \mathbf{y})$ is calculated by the above mentioned minimization.

From Lemma 4.1 the following theorem is derived.

Theorem 4.1 *If the submanifold $S = \{P(Z)|P(t(Z)) = p^{t*}\}$ is m-flat, the posterior probability of $P(Z)$ given the information of $P(t(Z)) = p^{t*}$ is given by the e-projection of $P(Z)$ on S. Especially, if p^{t*} is the point mass in $t(Z) = t$, the ordinary posterior probability $P(Z|t(Z) = t)$ is given by the e-projection.*

Since the submanifold $S = \{P(\mathbf{U}, \mathbf{X}, \mathbf{Y}, \mathbf{E})|P(\mathbf{E} = \mathbf{e}) = 1\}$ is m-flat, the posterior probability $P(\mathbf{U}, \mathbf{X}, \mathbf{Y}|\mathbf{E} = \mathbf{e})$ is given by the e-projection from an initial probability $P(\mathbf{U}, \mathbf{X}, \mathbf{Y}, \mathbf{E})$ to the submanifold S. Let $P_1 \to_S P_2$ (or $P_1 \to_{P(\mathbf{E})} P_2$) denote that the probability P_2 is the e-projection of a probability P_1 on the submanifold S.

Lemma 4.2 *Let S, S_1 and S_2 be the submanifold $\{P(Z)|P(t_1(Z)) = p^{t_1} \wedge P(t_2(Z)) = p^{t_2}\}$, $\{P(Z)|P(t_1(Z)) = p^{t_1}\}$ and $\{P(Z)|P(t_2(Z)) = p^{t_2}\}$ respectively. The iterative e-projections $P_0(Z) \to_{S_1} P_1(Z) \to_{S_2} P_2(Z) \to_{S_1} P_3(Z) \cdots$ converges at the e-projection of an initial probability $P_0(Z)$ on the submanifold S.*

Lemma 4.3 *Let S be the submanifold $\{P(\mathbf{E}, \mathbf{U}, \mathbf{X}, \mathbf{Y})|P(\mathbf{E} = \mathbf{e}) = P_e^*\}$. If*

$$P(\mathbf{E}, \mathbf{U}, \mathbf{X}, \mathbf{Y}) = \frac{P(\mathbf{E}, \mathbf{U})P(\mathbf{U}, \mathbf{X})P(\mathbf{X}, \mathbf{Y})}{P(\mathbf{U})P(\mathbf{X})}, \tag{4}$$

the following iterative procedure converges at the marginal probability $P(\mathbf{Y})$ of the e-projection of an initial probability $P_0(\mathbf{E}, \mathbf{U}, \mathbf{X}, \mathbf{Y})$ on the submanifold S.

The iterative procedure is constructed by e-projections and marginal operations as follows:

$$P_0(\mathbf{E}, \mathbf{U}) \to_{P(\mathbf{E})} P_1(\mathbf{E}, \mathbf{U}) \Rightarrow P_1(\mathbf{U}),$$

$$P_0(\mathbf{U}, \mathbf{X}) \to_{P_1(\mathbf{U})} P_1(\mathbf{U}, \mathbf{X}) \Rightarrow P_1(\mathbf{X}),$$

$$P_0(\mathbf{X}, \mathbf{Y}) \to_{P_1(\mathbf{X})} P_1((\mathbf{X}, \mathbf{Y}) \Rightarrow P_1(\mathbf{Y}),$$

where $\Rightarrow$ is a marginal operation.

Theorem 4.1 shows that the posterior probability given a condition is calculated by the e-projection of a prior probability on the submanifold satisfying the given condition. Lemma 4.2 shows that if multi-conditions are given then the posterior probability is calculated be repeating e-projection to the each condition. Lemma 4.3 shows that if we can assume conditional independence in a given model then the posterior marginal probability is calculated by the e-projection with respect to the marginal probability of a conditional independent block. From Theorem 4.1, Lemma 4.2 and Lemma 4.3, we obtain the following theorem.

Theorem 4.2 *If any collision does not occur in the main-procedure, the procedure calculate the exact posterior marginal probabilities for Problem 2.1.*

The theorem guarantees justice of the proposed algorithm.

5 The BP algorithm and the proposed algorithm

We shall compare the BP algorithm with the proposed algorithm.

First, we shall compare the applicable probability models of the algorithms. The primitive BP algorithm can not be applied to all BN. The application of the BP algorithm is restricted to polytrees, which is a subset of BN. The proposed algorithm can be applied to any log linear model. The applicable probability model class of the proposed algorithm is wider than that of the BP algorithm, since the log linear model class includes any model presented by BN.

In the probability model that is repesented by the BN in Fig. 1, the posterior marginal probability $P(\mathbf{X}|\mathbf{e}^+, \mathbf{e}^-)$ can be calculated by the BP algorithm. The marginal probability can be also given by the proposed algorithm. In BN, conditionally independence holds between the random variables whose nodes are not connected together by any arc.

The propagation π of the BP algorithm is interpreted as the procedure combining e-projection and marginal operation that is the same as that given in Lemma 4.3 as follows.

$$P_o(\mathbf{U}, \mathbf{X}) \to_{P_1(\mathbf{U}|\mathbf{E}^+)} P_1(\mathbf{U}, \mathbf{X}) \Rightarrow P_1(\mathbf{X}). \tag{5}$$

The inverse propagation λ is interpreted as the function of the probability $P_1(\mathbf{X})$ valued $P_2(\mathbf{X})$.

$$f(P_1(\mathbf{X})) = \alpha\lambda(\mathbf{X})P_1(\mathbf{X}) = P_2(\mathbf{X}). \tag{6}$$

The function is interpreted as the procedure in Lemma 4.3 as follows.

$$P_o(\mathbf{X}, \mathbf{Y}) \to_{P_1(\mathbf{X})} P_1(\mathbf{X}, \mathbf{Y}), \tag{7}$$

$$P_o(\mathbf{X},\mathbf{Y}) \to_{P_1(\mathbf{Y}|\mathbf{e}^-)} P_2(\mathbf{X},\mathbf{Y}) \Rightarrow P_2(\mathbf{X}) \quad (8)$$

Finally, the objective marginal probability is given by $\pi(\mathbf{X})$ and $\lambda(\mathbf{X})$ as follows.

$$P(\mathbf{X}|\mathbf{e}^+,\mathbf{e}^-) = \alpha\lambda(\mathbf{X})\pi(\mathbf{X}) = P_2(\mathbf{X}). \quad (9)$$

Thus Theorem 4.1 and Lemma 4.3 guarantee that the correct posterior marginal probability is calculated by the BP algorithm.

Pearl proved that the BP algorithm works exactly in polytrees[10]. So, this is considered as another proof of the justice of the BP algorithm.

If the proposed algorithm is applied to the model represented by a polytree, there occurs no collision in the calculation process. The probability model class for which the proposed algorithm guarantees exact answer is wider than that of the primitive BP algorithm.

6 Application to the turbo decoding

Although the BN representing compound codes, which include the original turbo codes, have loops, the primitive procedure of BP can not be directly used in the turbo decoding. In the BN that has loops, the BN is broken into tractable subnetworks that have no loops, and the BP algorithm is reciprocally applied to each subnetwork.

The justice of the BP algorithm is guaranteed by the assumption of conditionally independence of the random variables in a BN. The compound BP algorithm such as the turbo decoding cannot calculate correct posterior marginal probability, because the compound BP algorithm does not guarautee the e-projection.

The proposed algorithm can be directly applied to the probability model of the compound codes such as turbo codes. The algorithm is not necessary to break the model, i.e., the BN, into tractable submodels, i.e., subnetworks. If a collision does not occur in the calculation process, the results given by the proposed algorithm are exact. In the case that collisions occur, if the impact of the collision is not so big, the result is almost exact answer. The impact of collision means the difference between the new propagated marginal probability and the probability that have been already calculated by another propagation.

7 Conclusion

We investigated the calculation of posterior probability from the viewpoint of differential geometry. The e-projection to the submanifold representing a given codition attains the exact posterior probability. We have proposed the iteralive procedure that calculates an exact posterior marginal probabilities. We have also shown the BP algorithm is interpreted as the equivalent procedure of the proposed iterative procedure in the BN that dose not have loops. Although the turbo decoding algorithm, i.e., the compound BP algorithm applied to the BN including loops dose not guarantee an exact posterior probability, the proposed procedure guarantees that if there is no collision.

References

[1] S. Amari, *Differential-geometry in statistical inference*, Springer lecture notes in statistics vol.28, Springer, 1985.

[2] C. Berrou, A. Glavieux and P. Thitimajshima, *Near Shannon limit error-correcting coding and decoding*, in Proc. IEEE Int. Conf. Commun., 1993.

[3] J. Hagenauer, E. Offer and L. Papke, *Iterative decoding of binary block and convolutional codes*, IEEE Trans. IT, Vol.42, 1996.

[4] C. Heegard and S.B. Wicker, *Turbo coding* Kluwer Academic Publishers, 1999.

[5] F.R. Kschischang and B.J. Fey, *Iterative decoding of compound codes by probability propagation in graphical models*, IEEE J. Sel. Areas Commun., Vol.16 No.2, 1998.

[6] S.L. Lauritzen, D.J. Spiegelhalter, *Local computations with probabilities on graphical structures and their application to expert system*, J. Roy. Statist. Soc., Ser. B, Vol. 50, 1988.

[7] R.J. McElice, D.J.C. MacKay and J. Cheng, *Turbo decoding as an instance of Pearl's "Belief Propagation"*, IEEE J. Sel. Areas Commun., Vol.16 No.2, 1998.

[8] D.J.C. MacKay, *Good Error-Correcting Codes Based on Very Sparse Matrices*, IEEE IT., Vol.45 No.2, 1999.

[9] M.K. Murray, J.W. Rice, *Differential geometry and statistics* Chapman Hall, 1993.

[10] J. Pearl, *Probabilistic reasoning in intelligent systems* Morgan Kaufmann, 1988.

[11] M. Suzuki, T. Matsushima, S. Hirasawa, *On reasoning model and method for uncertainty*, Trans. of Information Processing Society of Japan, 2000.

[12] M. W. Birch, *Maximum likelihood in three-way contingency tables*, J. Roy. Statist. Soc. B, 25, 220-233, 1963.

[13] J. N. Darroch, *Interactions in multi-factor contingency tables*, J. Roy. Statist. Soc. B, 24, 251-263, 1962.

The 23rd Symposium on Information Theory
and Its Applications (SITA2000)
Aso, Kumamoto, Japan, Oct. 10 - 13, 2000

尤度比検定を用いた木符号の復号法について
A decoding algorithm using likelihood ratio testing for Tree Codes

新家 稔央 * 松嶋 敏泰 † 平澤 茂一 †
Toshihiro NIINOMI Toshiyasu MATSUSHIMA Shigeichi HIRASAWA

Abstract– The generalized Viterbi algorithm(GVA) is known as an efficient decoding method for tree codes, which employs fixed list decoder in the mean process. In this paper, we newly propose the algorithm, which does not use fixed size list decoder but variable size decoder with the decision criterion of likelihood testing.

Keywords– リスト復号法, 一般化 Viterbi 復号法, ランダム符号化, 判定基準

1 はじめに

Forney[6] は, 最尤復号の一般化について, はじめて理論的解析を行った. [6] では, 判定基準 (decision critera)[6](11) 式を設け, そのしきい値 T の値により, 消失判定やリストサイズを確率変数にとるリスト復号器 (variable size list decoder 以下, VLD と記す.) となることを示している. すなわち, 消失判定と VLD が, 表裏一体の復号法であることを指摘した.

一方, Anderson[2] によれば, 木符号の復号には, 幅優先 (横形探索 width first), 深さ優先 (縦型探索 depth first), メトリック優先 (metric first) の方法に大別することが可能である. 幅優先のアルゴリズムには, Viterbi アルゴリズム (VA) や一般化 Vitebi アルゴリズム (GVA) があげられるが, これらのアルゴリズムは「パス選択を, 同じ長さの受信系列に対する尤度の比較によって行い, パスの候補を絞り込んでいく」点で共通する[1].

したがって, 幅優先のアルゴリズムでは, パス選択において, Forney[6] の意味で一般化が可能である[2]. たとえば, VA は, 幅優先のアルゴリズムだが, ただ1本の最尤パスを選択するアルゴリズムである. これを一般化し, 消失判定が行えるようにできれば, 帰還通信路と組み合わせて判定帰還方式が構成できる. [3] では, 各ノードごとに行われるパス選択の際に, 判定基準として簡略的な尤度比検定を用い, その結果をサバイバにラベルづけすることで, このことを実現している.

同じように, 幅優先のアルゴリズムに尤度比検定を取り入れ, ノードごとに VLD とる復号が自然に考えられる. また, 従来, 固定サイズのリスト復号の代表的応用例として, GVA がノードごとに複数本のパスを選択を行う例が有名である.

そこで, 本研究では, GVA が各々のノードでサバイバを複数残す際, VLD を用いてできる新しい復号アルゴリズムを提案し, その性質を考察することを目的とする. 前述したように, VLD には判定基準が必要だが, 本研究では, 判定帰還 [3] で用いられた尤度比検定のしきい値を1より小さくして用い, VLD として用いる意味での一般化を行う. なお, 通信路容量 C の離散的無記憶通信路を仮定する.

2 提案アルゴリズム

提案するアルゴリズムでは, GVA[1] と同様, 設定された復号制約長 L に対し, q^{L-1} 個の各ノードごとに複数本のサバイバを残すが, 判定基準を設けて VLD [3] を行うので, その本数は GVA と異なり固定でない. また, 勝ち残ったサバイバの集合をリストと呼ぶ.

2.1 準備

以下では入力アルファベット $A = \{0, 1, \cdots, a-1\}$, 出力アルファベット $B = \{0, 1, \cdots, b-1\}$ の離散無記憶通信路 $P = \{P_{ij}, j \in A, i \in B\}$ を仮定し, 送信される情報記号系列 (メッセージ) を $\mathbf{u}^N$ で表す. $\mathbf{u}^N$ は, 符号器へ入力される q 元アルファベット $\mathcal{U} = \{0, 1, \cdots, q-1\}$ からなる長さ N の系列である. したがって, ある情報記号系列 $\mathbf{u}_i{}^N$ は, $\mathbf{u}_i{}^N = u_{i,1}u_{i,2}\cdots\cdots u_{i,t}\cdots u_{i,N}$, ただし, $u_{i,t} \in \mathcal{U}$, $t = 1, 2, \cdots, N$, $i = 1, 2, \cdots, q^N$ で示される. 木符号は, q 進木で示すことができ, ある情報記号系列はルートから伸びる1本のパスに対応する. すなわち, $\mathbf{u}_i{}^N$ の第 t ブランチは, $u_{i,t}$ に対して符号化が行なわれている. 符号化に用いる1ブランチあたりの通信路の入力アルファベットの数を v とすると, レート R は, $R = \frac{1}{v}\ln q$ と定義できる.

なお, パス $\mathbf{u}_i{}^N$ の第1ブランチから第 n ブランチまでの部分系列を $\mathbf{u}_i{}^n$ で記す. また, 第1ブランチから第 n ブランチまでのパス $\mathbf{u}_i{}^n$ を情報系列とする符号系列および受信系列を, それぞれ, $\mathbf{x}_i^{vn}$ および $\mathbf{y}^{vn}$ と表し, その第 t ブランチに対する符号系列および受信系列を, それぞれ, $\mathbf{x}_{i,t}^{v}$ および $\mathbf{y}_t^v$ で記すことにする. さらに, チェックテイルの長さを T ブランチで表す.

2.2 アルゴリズム

GVA と同様, 以下に示すアルゴリズムでは, (1)〜(3) の再帰手続きと, チェックテイルにおけるパス選択とに分かれる. 本稿で提案するアルゴリズムを [1] と同様, 次のように示す.

* 神奈川工科大学情報ネットワーク工学科, 厚木市下荻野 1030. Dept. of Information Network Engineering, Kanagawa Institute of Tech., 1030 Shimo-Ogino, Atsugi-shi, Kanagawa, Japan. niinomi@ele.kanagawa-it.ac.jp

† 早稲田大学理工学部経営システム工学科, 新宿区大久保 3-4-1 School of Science and Engineering, Waseda University, 3-4-1 Ohkubo Shinjyuku-ku, Tokyo, Japan

[1] 一方, 残りの2者に対しては, 異なった長さの受信系列に対して比較を行う. 例えば, スタックアルゴリズムなどは, 異なった長さのパスに対する尤度を正規化して比較している. これはメトリック優先のアルゴリズムに属する.

[2] メトリック優先の復号法であるスタックアルゴリズムをベースとして消失判定を行う方式も提案されている [8] が, 同じ長さの受信系列に対するパス比較ではないため, 尤度比の検定を判定基準として用いていない.

[3] ブロック符号において, リスト復号誤り確率と平均リストサイズを同時に最小にする最適な判定基準は, [6] に記された判定基準である. この判定基準は, 判定に要する計算量を増加させる意味で, 木符号や畳込み符号との整合がよくない. このことは, VA を用いた判定帰還に対する論文 [4] でも指摘されていた. 近年, この判定基準と同等の判定を用いる判定帰還 [5] が研究されているが, 本稿では, 判定に要する計算量を増やさない意味で整合のとれた [3] の判定基準を一般化して用いる.

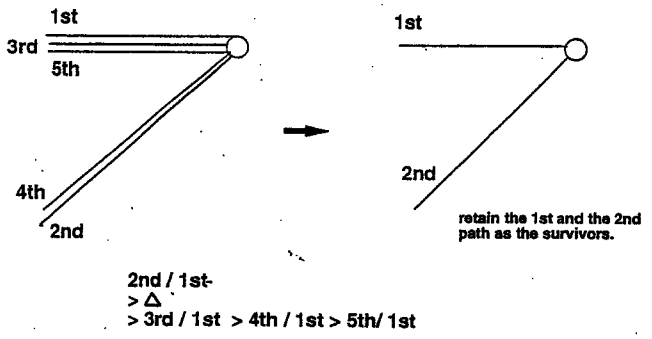

図 1: リストの選択

[再帰手続き] レベル L まで, すべてのパスを伸長する. そして, レベル $n(L \leq n \leq N)$ に対して以下の手続きを繰り返す.

（1）初期条件

レベル $n-1$ において, q^{L-1} 個の状態ごとに, すべてのパスを保持し, これをリストの初期条件とする. 要素は長さ $n-1$ ブランチのパスである.

（2）パス伸長

q^{L-1} 個のリストの各々のサバイバに対して, 1ブランチの伸長を行なう. すなわち, 情報記号列 $\mathbf{u}^{n-1}$ に対応するパスがサバイバとして残っているとしたとき, q 本のパスを伸長し, $\mathbf{u}^n = \mathbf{u}^{n-1}u$, $u \in \mathcal{U}$ に対するパスの尤度を計算する.

（3）VLD によるパス選択

レベル n の q^{L-1} の各々のノードに合流するパスの中で k 番めに尤度の大きいパスを $\mathbf{u}^n_{(k)}$ [4] で表す. 合流するすべてのパスに対して, 以下の判定基準[5] により, サバイバを決定する. すなわち, m をこのノードに合流するパスとしたとき,

$$\frac{Pr(\mathbf{y}^{vn}|\mathbf{x}^{vn}_m)}{Pr(\mathbf{y}^{vn}|\mathbf{x}^{vn}_{(1)})} \geq \Delta, \quad \Delta < 1 \qquad (1)$$

であれば, サバイバとする. そうでなければ, 捨てる. 合流するすべてのパスに対して, (1) を繰返して選ばれたサバイバの集合を, このノードのリストとする.

[チェックテイルにおけるパス選択]

長さ $L-1$ の既知シンボルの列 $\mathbf{u}^{L-1} \in \mathcal{U}^{L-1}$ により, 再び上記（1）～（3）を用い, レベル $N+L-1$ におけるただ1つのノードで得られるリストに含まれたサバイバのみに絞る. さらに, $T-(L-1)$ 個の既知シンボルによる符号語に対して尤度を計算し, レベル $N+T$ において, この中から最も尤度の大きいパスを最終的に選択する.

[4] 本稿では, レベル n のノードにおいて, 情報記号列 $\mathbf{u}_k{}^N$ に対する尤度を $Pr(\mathbf{y}^{vn}|\mathbf{x}^{vn}_k)$ と表すのに対し, 尤度が k 番目に大きいパスの尤度を $Pr(\mathbf{y}^{vn}|\mathbf{x}^{vn}_{(k)})$ と使い分ける. すなわち, $(\cdot)$ はあるノードでのパスの尤度に対する順位を表す添字である.

[5] [3] では, VA の各ノードごとに, $\Delta > 1$ として, しきい値と比べ, 最大尤度のパスに *Accept* か *Reject* のラベルづけを行い, すべてのサバイバのラベルが, *Reject* になったとき, 消失と判定する. また, [3] のアルゴリズムが, トレリス全体のパスの中から, 必ず最大尤度と2番目のパスの比較を行うのに対し, 提案アルゴリズムでは木符号に対する復号アルゴリズムである点が異なる.

3 ランダム符号化による評価

3.1 準備

実際に送信された情報記号列が $\mathbf{u}_0{}^N$ のとき, 次の確率を定義する.

$P_e(\mathbf{u}_0{}^N) = [\mathbf{u}_0{}^N$ が復号されない確率]

$P_{ea}(\mathbf{u}_0{}^N) = [$レベル $N+L-1$ で作られるリストから $\mathbf{u}_0{}^N$ が外れる確率]

$P_{eb}(\mathbf{u}_0{}^N) = [$ レベル $N+T$ において, リストの中からパスを1本に絞ったとき, $\mathbf{u}_0{}^N$ が選ばれない確率]

また, これらの確率を $P_*(\mathbf{u}^N)$ で表したとき, すべてのメッセージに対する平均を

$$P_* \stackrel{\text{def}}{=} \frac{1}{q^N}\sum_{\mathbf{u}^N} P_*(\mathbf{u}^N)$$

と表すことにする. ただし, $* \in \{e, ea, eb\}$. さらに, Forney [6] に習い, 復号誤り確率を $Pr(E_1)$, レベル $N+T$ において実際に送信された情報記号列以外のパスがリストに含まれる確率を $Pr(E_2)$ とする[6] . $Pr(E_1)$ について, 以下の上界が成り立つ.

$$Pr(E_1) = P_e \leq P_{ea} + P_{eb}$$

ここで, P_{ea}, P_{eb} は,

$$P_{ea} \leq \frac{1}{q^N}\sum_{\mathbf{u}_0{}^N}\sum_{n=L}^{N+L-1} P_{ea,n}(\mathbf{u}_0{}^N),$$

$$P_{ea,n}(\mathbf{u}_0{}^N) = Pr\Big[\text{レベル } n \text{ のノードで, ある } i \neq 0 \text{ に対し, } \frac{Pr(\mathbf{y}^{vn}|\mathbf{x}^{vn}_0)}{Pr(\mathbf{y}^{vn}|\mathbf{x}^{vn}_i)} \leq \Delta\Big],$$

$$P_{eb} \leq \frac{1}{q^N}\sum_{\mathbf{u}_0{}^N} P_{eb}(\mathbf{u}_0{}^N),$$

$$P_{eb}(\mathbf{u}_0{}^N) \leq Pr\Big[\text{レベル } N+T \text{ のノードで, ある } i \neq 0 \text{ に対し, } \frac{Pr(\mathbf{y}^{v(N+T)}|\mathbf{x}^{v(N+T)}_i)}{Pr(\mathbf{y}^{v(N+T)}|\mathbf{x}^{v(N+T)}_0)} \geq 1\Big]$$

である. $P_{ea,n}(\mathbf{u}_0{}^N)$ は, レベル n でリスト復号誤りが起きる確率である. この確率は, 正しいパスとそれ以外のパスで最も尤度の大きいパスとの尤度比が, しきい値 Δ をこえない確率で上界できる. また, P_{eb} は, レベル $N+L-1$ で正しいパスが残っている場合に, レベル $N+T$ で復号誤りのおきる確率である[7].

$Pr(E_2)$ についても, 同じくユニオン上界が可能である. すなわち,

$$Pr(E_2) \leq \frac{1}{q^N}\sum_{\mathbf{u}_0{}^N}\sum_{n=L}^{N+L-1} P_{E2,n}(\mathbf{u}_0{}^N),$$

$$P_{E2,n}(\mathbf{u}_0{}^N) = Pr\Big[\text{レベル } n \text{ のノードで, ある } i \neq 0 \text{ に対し, } \frac{Pr(\mathbf{y}^{vn}|\mathbf{x}^{vn}_i)}{Pr(\mathbf{y}^{vn}|\mathbf{x}^{vn}_0)} > \Delta\Big],$$

[6] $Pr(E_1) = P_e$ ゆえ, 2重の表記が不適当にも感じるが, $Pr(E_1)$ と $Pr(E_2)$ がしきい値 Δ によって, トレードオフの関係があることを強調したいため, このような表記を用いた. なお, [6] では, $Pr(E_2)$ をリストサイズの平均値と定義しており, 本稿と異なっている点に注意したい.

[7] したがって, P_{eb} は, 正しいパスの尤度がレベル $N+T$ において, 比較されるどれかのパスの尤度より小さくなる確率で上界できる.

が成り立つ. $P_{E2,n}(\mathrm{u_0}^N)$ は, レベル n のあるノードで, 実際に送信されたパスが生き残っているとき, このノードで選ばれたリストに 1 本以上の誤りパスが生き残る確率である.

3.2 主要な結果

ランダム符号化を用いた評価結果を以下の補題に示す. はじめに, 補題 1 および補題 2 で, アンサンブルに対する P_{ea} と $Pr(E_2)$ の平均の上界を示す. 次に, チェックテイルの長さをある程度とれば, $Pr(E_1)$ は, P_{ea} が支配することを [1] の結果を用いて示す. 最後にこれらより, 提案したアルゴリズムが達成し得る誤り指数の下界を求めて定理に示す. なお, $\mathbf{q} = \{q_0, q_1, \cdots, q_{a-1}\}$ を通信路シンボルの入力確率分布とする.

[補題 1] ランダムな木符号を用いることで, P_{ea} のアンサンブルに対する平均 $\mathcal{E}P_{ea}$ が次式で上界される.

$$\begin{aligned}
&\mathcal{E}P_{ea} \le \frac{Ne^{vR\rho_1}}{1-e^{-v\epsilon_1}} \\
&\cdot \exp\left\{-Lv\left[E_o(\sigma_1,\rho_1,\mathbf{q}) - \sigma_1\frac{\ln\Delta}{Lv}\right]\right\}, \\
&0 \le \rho_1 \le 1, \quad \sigma_1 \ge 0, \\
&\epsilon_1 = E_o(\sigma_1,\rho_1,\mathbf{q}) - \rho_1 R > 0, \\
&E_o(\sigma_1,\rho_1,\mathbf{q}) \\
&= -\ln\left[\sum_{j\in B}\left(\sum_{i\in A} q_i P_{ji}^{1-\sigma_1}\right)\left(\sum_{k\in A} q_k P_{jk}^{\sigma_1/\rho_1}\right)^{\rho_1}\right].
\end{aligned} \tag{2}$$

(証明省略)

[補題 2] ランダムな木符号を用いることで,$Pr(E_2)$ のアンサンブルに対する平均 $\mathcal{E}Pr(E_2)$ が次式で上界される.

$$\begin{aligned}
&\mathcal{E}Pr(E_2) \le \frac{Ne^{vR\rho_2}}{1-e^{-v\epsilon_2}} \\
&\cdot \exp\left\{-Lv\left[E_o(\sigma_2,\rho_2,\mathbf{q}) + \sigma_2\frac{\ln\Delta}{Lv}\right]\right\}, \\
&0 \le \rho_2 \le 1, \quad \sigma_2 \ge 0, \\
&\epsilon_2 = E_o(\sigma_2,\rho_2,\mathbf{q}) - \rho_2 R > 0.
\end{aligned} \tag{3}$$

(証明省略).

次に, チェックテイルをある程度とれば, P_{eb} は P_{ea} に対して無視できることを示そう. はじめに, P_{eb} に注目する. 補題 3 に [1] で得られた結果を示す.

[補題 3] ランダムな木符号を用いることで, P_{eb} のアンサンブルに対する平均 $\mathcal{E}P_{eb}$ が次式で上界される.

$$\begin{aligned}
&\mathcal{E}P_{eb} \le \frac{e^{vR\rho'}}{1-e^{-v\epsilon'}} \\
&\cdot \exp\left[-v(T+1)E_o(\rho',\mathbf{q})\right], \\
&0 \le \rho' \le 1, \ \epsilon' = E_o(\rho',\mathbf{q}) - \rho' R > 0, \\
&E_o(\rho',\mathbf{q}) \\
&= -\ln\left[\sum_{k\in B}\left(\sum_{j\in A} q_j P_{kj}^{\frac{1}{1+\rho'}}\right)^{1+\rho'}\right].
\end{aligned} \tag{4}$$

(証明)[1]p.875, A-9 参照.

そこで, 補題 2, 補題 3 より, チェックテイルの長さを

$$T \ge \left[\frac{E_o(\sigma_1,\rho_1,\mathbf{q}) - \sigma_1\frac{\ln\Delta}{Lv}}{E_o(\rho',\mathbf{q})}\right]L - 1 \tag{5}$$

としてやれば, 全体の復号誤り確率 $Pr(E_2)$ は, 漸近的に P_{ea} の指数部 $E_o(\sigma_1,\rho_1,\mathbf{q}) - \sigma_1\frac{\ln\Delta}{Lv}$ によって支配される.

また, 次の補題が [4] と同様に導かれる.

[補題 4] 次式を同時に満たす木符号が存在する.

$$\begin{aligned}
Pr(E_1) &\le 2\mathcal{E}Pr(E_1), \\
Pr(E_2) &\le 2\mathcal{E}Pr(E_2).
\end{aligned} \tag{6}$$

(証明)[14] 参照

以上より, 提案アルゴリズムが達成し得る誤り指数

$$-\frac{1}{vL}\ln Pr(E_1), \ L\to\infty$$

の下界を求めてみよう.

このアルゴリズムでは, しきい値が小さいほど, 復号誤り確率の指数部を大きくできる. ところが, しきい値を無造作に小さくとったのでは, チェックテイルに至るまでに莫大なサバイバが生き残り, 結果として, チェックテイルで 1 本のパスを選択するための比較回数が全体の計算量を支配してしまう. すなわち, アルゴリズム前半の再帰手続きが意味を失う. そこで, L を大きくとれば, レベル $N+T$ で作られるリストに誤りパスが含まれる確率を任意に小さくできることを条件に, しきい値を設定をする. すなわち, $-\frac{1}{vL}\ln Pr(E_2) \to 0 (L\to\infty)$ であるようにしきい値を定め, $\frac{\ln\Delta}{Lv}$ の下界を求める. その上で, σ_2, ρ_2 を最適化する. したがって, 提案アルゴリズムが漸近的に達成可能な誤り指数の下界 $e_1^{new}(R)$ 次式のようになる.

$$\begin{aligned}
e_1^{new}(R) = \max_{\mathbf{q},\sigma_1,\rho_1\in\mathcal{D}_1}\Bigg\{&E_o(\sigma_1,\rho_1,\mathbf{q}) \\
&- \sigma_1\cdot\min_{\mathbf{q},\sigma_2,\rho_2\in\mathcal{D}_2}\left[-\frac{E_o(\sigma_2,\rho_2,\mathbf{q})}{\sigma_2}\right]\Bigg\},
\end{aligned} \tag{7}$$

$$\begin{aligned}
\mathcal{D}_1 = \{&0 \le \rho_1 \le 1, \ \sigma_1 \ge 0, \\
&\epsilon_1 = E_o(\sigma_1,\rho_1,\mathbf{q}) - \rho_1 R > 0\},
\end{aligned} \tag{8}$$

$$\begin{aligned}
\mathcal{D}_2 = \{&0 \le \rho_2 \le 1, \ \sigma_2 \ge 0, \\
&\epsilon_2 = E_o(\sigma_2,\rho_2,\mathbf{q}) - \rho_2 R > 0\}.
\end{aligned} \tag{9}$$

そこで

$$e_F(R) \stackrel{\text{def}}{=} \max_{\mathbf{q},\sigma_2,\rho_2\in\mathcal{D}_2}\frac{E_o(\sigma_2,\rho_2,\mathbf{q})}{\sigma_2} \tag{10}$$

と定義する[8]. $E_o(\sigma_2,\rho_2,\mathbf{q})$ は, σ_2 に対して上に凸な関数であるので, $\frac{E_o(\sigma_2,\rho_2,\mathbf{q})}{\sigma_2}$ の上界は, $\rho_2 = \nu\sigma_2$ とおけば,

$$e_F(R) = \lim_{\sigma_2\to 0}\frac{E_o(\sigma_2,\nu\sigma_2,\mathbf{q})}{\sigma_2}, \tag{11}$$

$$\frac{E_o(\sigma_2,\nu\sigma_2,\mathbf{q})}{\sigma_2} - \nu R > 0 \tag{12}$$

で得られる. そこで, ロピタルの定理より,

$$\lim_{\sigma_2\to 0}\frac{E_o(\sigma_2,\nu\sigma_2,\mathbf{q})}{\sigma_2} \tag{13}$$

$$= \frac{\partial}{\partial\sigma_x}E_o(\sigma_2,\nu\sigma_2,\mathbf{q})\ \Big|_{\sigma_x=0} \tag{14}$$

[8] $e_F(R)$ は, Forney のブロック符号における帰還誤り指数と連接構造をなす誤り指数 $e_F(R)$, [4]p.568,(11) 式と一致する.

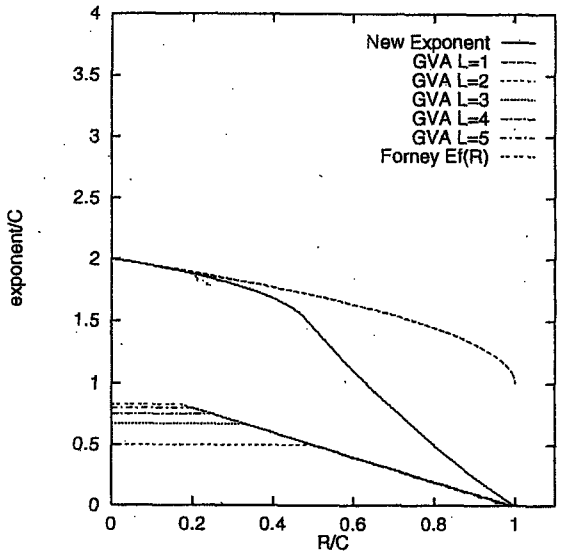

図 2: VNC における誤り指数

を用いて次の定理を得る.

[定理] 提案アルゴリズムに対する誤り指数の下界 $e_1^{new}(R)$ は,

$$e_1^{new}(R) = \max_{\mathbf{q},\sigma_1,\rho_1 \in \mathcal{D}_1} \left\{ E_o(\sigma_1,\rho_1,\mathbf{q}) + \sigma_1 \cdot e_F(R) \right\},$$
$$e_F(R) = \max_{\mathbf{q},\nu \in \mathcal{D}_3} E_{oF}(\nu,\mathbf{q}),$$
$$\mathcal{D}_3 = \{ E_{oF}(\nu,\mathbf{q}) - \nu R > 0, \quad \nu > 0 \},$$
$$E_{oF}(\nu,\mathbf{q}) = \sum_{k\in B}\sum_{j\in A} q_j P_{kj} \ln \left[\frac{P_{kj}^{1/\nu}}{\sum_{j\in A} q_j {P_{kj}}^{1/\nu}} \right]^{\nu} \tag{15}$$

で与えられる[9].

残念ながら, 一般の離散無記憶通信路においては, $e_1^{new}(R)$ をこれ以上, 簡単に表現することはできない. そこで, 強雑音通信路 (例えば [11][12] などを参照) において, $e_1^{new}(R)$ を計算する. $\mathbf{q}$ を最適化して得られる通信路容量 C の強雑音通信路を仮定すると,

$$E_o(\sigma_1,\rho_1,\mathbf{q}) = \sigma_1 \left[2 - \sigma_1 \left(1 + \frac{1}{\rho_1} \right) \right] C$$
$$E_{oF}(\nu,\mathbf{q}) = \left(2 - \frac{1}{\nu} \right) C$$
$$e_F(R) = \left(1 + \sqrt{1 - \frac{R}{C}} \right) C$$

であるので, これより, $e_1^{new}(R)$ を数値計算すると, 図 2 に示したようになる.

4 まとめ

一般化 Viterbi アルゴリズム (GVA) が固定されたリストサイズ (fixed sizelist decoder) のリスト復号器を用いるのに対し, 本稿ではリストサイズが確率変数であるリスト復号器 (variable size list decoder) を適用したときの新しい簡略的復号法を提案した. そして, その復号拘束長を大きくしたとき, 提案アルゴリズムの有効性を示す符号化定理を与えた.

謝辞 日頃ご指導いただく, 神奈川工科大学 石坂充弘教授に感謝いたします.

参考文献

[1] T.Hashimoto, "A list-type reduced-constraint generalization of the Viterbi algorithm", IEEE Trans. Inf.Theory vol.IT-33, no.6, pp.866-876, Nov.1987.

[2] J.B.Anderson and S.Mohan, "Sequential coding algorithms: A survey and cost analysis" IEEE Trans. Commun. Vol.COM-32, No.2, Feb.1984.

[3] H.Yamamoto and K.Itoh, " Vitebi decoding algorithm for convolutional codes with repeat request" IEEE Trans. Inf.Theory vol.IT-26, no.5, pp540-547, Sep.1980.

[4] T.Hashimoto, "On the error exponent of convolutionaly coded ARQ", IEEE Trans. Inf.Theory vol.IT-40, no.2, pp.567-575, Mar.1994.

[5] T.Hashimoto, "Composite scheme LR + Th for decoding with erasures and its effective equivalence to Forney's rule", IEEE Trans. Inf.Theory vol.45, pp.78-93, Jan.1999.

[6] G.D.Forney,Jr., "Exponential error bounds for erasure,list and decision feedback schemes", IEEE Trans.Inf.Theory vol.IT-14, no.2, pp.206-220, Mar.1968.

[7] G.D.Forney,Jr., "Convolutional codes II : Maximum likelihood decoding", Inf.Control. vol.25, no.3, pp.222-266, Jul.1974.

[8] A.Drukarev, "Hybrid ARQ error control using sequntial decoding", IEEE Trans. Inf.Theory vol.IT-23, no.4, pp.311-318, 1970.

[9] C.E.Shannon, R.G.Gallager, and E.R.Berlekamp, "Lower bounds to error probability for coding for discrete memoryless channels", Inf.Control, vol.19, no.1, pp.65-103, 1967.

[10] J.L.Massey and D.J.Costello, "Nonsystematic convolutional codes for sequential decoding in space applications", IEEE Trans. Commun.Technol. vol.COM-19, no.5, pp.806-813, Oct.1971.

[11] A.J.Viterbi and J.K.Omura, " Principles of communication and coding", NY:McGraw-Hill, 1979.

[12] R.G.Gallager, "Imformation theory and reliable communication", NY:Wiley, 1968.

[13] G.D.Forney,Jr., "Concatenated codes", MA:M.I.T., 1966.

[14] 新家稔央, 松嶋敏泰, 平澤茂一, "木符号におけるリスト復号法を用いた判定帰還方式について", 信学論 (A), vol.83, no.1, pp.67-82, 2000.

[9] $e_1^{new}(R)$ は, 文献 [4] において, p.571 Corollary 2 の $e_1(R,p)$ に一致する.

2001 年 ベイジアンネットチュートリアル
Tutorial on Bayesian Networks (BN2001)
Tokyo, Japan, July 29, 2001.

Turbo符号，LDPC符号の復号アルゴリズム
Posterior Probability Calculation and the Decoding Algorithms for Turbo codes and LDPC codes

松嶋 敏泰* Toshiyasu MATSUSHIMA
早稲田大学 理工学部 Waseda University

Abstract: 情報理論、符号理論の分野では，近年，Turbo 符号及び LDPC 符号が理論限界に非常に近い高い性能を保持していることが実験的に実証され，次世代の移動体通信の方式に採用されるなど注目を集めている．このような高性能が実現された背景には，これらの符号の潜在的性能の高さと共に，誤り率最小復号を近似的に実現する効率良い復号アルゴリズムの存在を見逃すことは出来ない．この復号アルゴリズムが Bayesian network の代表的確信度更新アルゴリズムである Belief Propagation と等価であることが最近明らかになった．これらの復号アルゴリズムと確信度更新アルゴリズムが処理している問題は事後確率の計算のと解釈され，その他の多くの分野でも重要な問題となっている．本稿ではこの事後確率計算の問題を，確率分布のグラフ表現と，そのグラフを用いた効率良い計算アルゴリズムの観点から整理し，誤り訂正符号とその復号問題に適用した場合を概説する．

1 はじめに

誤り訂正符号は情報化社会において情報伝送，蓄積の基盤技術として欠くことの出来ない技術となっている．誤り訂正符号の性能は復号誤り率，符号化率，計算量などで評価され，その代表的理論的限界には Shannon 限界がある．現在実用化されている符号の多くは Shannon 限界に遠く及んでいなかったが，近年，Shannon 限界に近づく符号として Turbo 符号 [3] [6] [7] や LDPC(Low Density Parity Check) 符号 [5] [16] が注目されている．これらの符号の符号自体の潜在的性能の高さは様々な研究から明らかになりつつあるが，それと共に，誤り率最小復号を近似的に実現する効率良い復号アルゴリズムについても注目が集まり，多くの研究成果がでてきている [11][14]．また，その概念は様々な符号の復号や復調，検出など伝送システム全般へと適用されつつある [10]．

誤り訂正符号の復号問題は受信系列から情報系列を推定する問題と考えることができ，その本質は受信系列を与えられたもとでの情報系列の事後確率の計算問題に帰着される．また，BN(Bayesian Network) を用いた知識表現の分野において，証拠 (evidence) が与えられたもとで各命題の確信度 (belief)[15][18] の更新を行うことは，ある確率変数の値が与えられたもとでのその他の確率変数の事後周辺確率を求めていることに他ならない．よって両者は共通の問題と考えることができ，近年 Turbo 復号アルゴリズムや LDPC 符号で用いられる sum-products アルゴリズムは BN の代表的推論アルゴリズムである BP(Belief Propagation) と等価であることが明らかになった．このような事後確率の計算問題は符号理論における復号問題や人工知能分野における不確実な知識を扱う問題のみならず，統計学をはじめ，統計力学の分野，学習理論の分野，パターン認識の分野など様々な分野において共通で重要な問題となっている．

本稿では，まず誤り訂正符号の復号問題を事後確率の計算問題として整理する．次に確率変数間の従属関係が部分的に存在する確率分布のグラフを用いた代表的表現法と，そのグラフ上で情報を伝搬させることにより効率的に事後確率を計算するアルゴリズムについて概観する．最後にこれらを踏まえて，誤り訂正符号の復号問題における，符号のグラフ表現とグラフ上の伝搬を用いた効率的で高性能な復号アルゴリズムについて説明する．

2 誤り訂正符号の復号と事後確率計算

誤り訂正符号 [8] を用いた通信，蓄積についてまず簡単に説明する．送信側ではまず送信したい情報系列

* 〒 169-8555 東京都新宿区大久保 3-4-1, tel&fax: 03-5286-3301, e-mail: toshi@matsu.mgmt.waseda.ac.jp

$\mathbf{u} = (u_1, \cdots, u_K)$, $u_k \in A$ から符号化（関数）C により符号語 $C(\mathbf{u})$ を生成し，通信路を通して受信側へ送る．受信側では送信側から送られた符号語が雑音などの影響により一部のシンボルが異なってしまった受信系列を受け取ることになり，この受信系列を元の符号語または情報系列にもどすことを復号と呼ぶ．

この分野で中心的に研究されている符号はシンボル長 K の情報系列 $\mathbf{u}$ に $K \times N$ 生成行列 G をかけて[1]シンボル長 N の符号語を生成する線形符号のクラスである．

$$C(\mathbf{u}) = \mathbf{u}G.$$

生成行列の左側 $K \times K$ が単位行列であった場合，符号語は $(\mathbf{u}, \mathbf{x})$ となる．$\mathbf{x}$ をパリティと呼び，このような符号を組織符号と呼ぶ．

生成行列 G に対して $GH^T = 0$ を満たす行列 H を検査行列と呼び，すべての符号語が以下の性質を満たすことは符号語の生成過程より明らかであろう．

$$C(\mathbf{u})H^T = 0.$$

通信路で誤りが生じて符号語と異なってしまった場合の受信系列 $\mathbf{y}$ は，検査行列 H をかけても 0 にはならず，誤りが生じたことが受信側で検出される．この積をシンドロームと呼び，これを手がかりにある範囲内の誤りに対しては，ある種の方程式を解くことによって，誤りが生じたシンボルを求めることができる．このような代数的復号法が現在実用的には多く用いられている．この代数的復号法は計算量が少なく，訂正可能な領域が理論的に保証されているなどの利点を持っている．

結局この復号の問題は，受信系列を $\mathbf{y}$ としたもとで，受信系列 $\mathbf{y}$ から符号語 $C(\mathbf{u})$ または情報系列 $\mathbf{u}$ を推定する問題に帰着される．推定は情報系列全体 $\mathbf{u}$ をブロックとして推定する場合と各情報シンボル u_k を推定する場合に大別される．一般に推定の評価には誤り率が用いられ，前者に対するものをブロック誤り率，後者に対するものをシンボル誤り率と呼んでいる．

ブロック誤り率を最小化する推定（復号）法は受信系列を与えられたもとでの事後確率 $P(\mathbf{u}|\mathbf{y})$ を最大化する情報系列 $\hat{\mathbf{u}}$ を推定値とすることになる．これは $\mathbf{u}$ の事前確率が一様な場合は，尤度 $P(\mathbf{y}|\mathbf{u})$ を最大化する情報系列と一致し，通常これを最尤復号と呼んでいる．一方，シンボル誤り率を最小化する推定法は事後確率 $P(u_k|\mathbf{y})$ を最大化する情報シンボル $\hat{u}_k$ を推定値とすることになる．通常これを MAP(Maximum a posteriori probability) 復号[2] または MPM(Maximum posterior marginal) 復号などと呼んでいる．

$\mathbf{u}$ を推定する場合も u_k を推定する場合もそれぞれの確率変数の事後確率を計算し，その最大値を探索するという意味で基本的には同じ手続きで復号は行われるが，計算の困難な部分は両者で異なっている．一般に情報系列全体の事後確率 $P(\mathbf{u}|\mathbf{y})$ は比較的容易に計算可能であるが，最大値をとる $\hat{\mathbf{u}}$ を $|A|^K$ の候補から探索する問題は多くの計算量が必要とされる．この状況は尤度最大化の問題でも同様で，通信路復号化の多くの研究はこの最適な最尤復号に近い誤り率を，少ない計算量で実現する事に向けられてきた．例えば代数的復号だけでは最尤復号と比べ誤り率が高くなってしまうので，代数的復号を繰り返して用いることにより最尤復号またはそれに近い復号を実現するアルゴリズムが多く研究されている．

逆に，個々の情報シンボル u_k を推定する場合において，最大値を見つけることは $|A| - 1$ の比較で容易であるが，各情報シンボルの事後確率 $P(u_k|\mathbf{y})$ を計算することは，事後結合確率 $P(\mathbf{u}|\mathbf{y})$ から周辺確率を計算することに対応し，一般には $|A|^{K-1}$ 回の加算が必要となり指数オーダの計算量となる．この事後周辺確率計算を，少ない計算量で近似計算することに成功したある種の反復アルゴリズムが，Turbo 復号や LDPC 符号における sum-products アルゴリズムである．

先に述べたように，事後確率の値を求めることがその問題の主要な課題となっている問題は様々な分野で見ることができる．BN をはじめとする確率推論の分野おいて，証拠が与えられもとで各命題の確信度の更新を行うことは主要な問題であり，この問題はある確率変数の値が与えられたもとでのその他の確率変数の事後周辺確率を求めている上記の問題と同値の問題となっている．

3 確率変数間の相互関係のグラフによる表現

事後周辺確率の計算は一般に多くの計算量を必要とするが，確率変数間の相互の関連が部分的にのみ存在する確率分布においてはその計算が容易になる可能性がある．相互の関連が部分的とは，系全体の確率構造を表す結合確率が，確率変数の部分集合の関数の積で表される場合を指す．その関数としては，例えば条件付き確率やポテンシャル関数が用いられている．

このような条件が当てはまる確率分布として典型的なのはマルコフモデルであろう．隠れマルコフモデルなども含めたマルコフモデルは時系列解析，制御理論，信号処理，音声処理など様々な工学分野で用いられている．1 次元の広がりの関連を扱ったマルコフモデルを一般化

[1] A は有限集合なのでこれらの演算は有限代数上の演算となる．

[2] 正確には情報シンボルに対する最大事後確率復号と呼び，情報系列（ブロック）に対する最大事後確率復号と区別すべきであろうがここでは慣例に従う．

し，ある要素（確率変数）の確率が，近傍の要素が与えられたもとでの条件付き確率のみで決まるマルコフランダム場 (MRF) は，統計力学や画像処理の分野で使われている．人工知能の分野の不確実な知識処理の問題でも，確率事象や命題間の関係に条件付き独立が仮定される場合は多い．

これらの部分的に確率変数が関連を持つ分布の確率構造を表現するには，確率変数を節点に対応させ，関連のある確率変数の節点間を枝で結んだグラフを用いると分かりやすく表現できる．用いられるグラフ表現は分野によって様々であるが，まず無向グラフと有向グラフに大別される．前者としてはMRF[11]，moralグラフ [13] 等があり，後者として代表的なのがBNである．BNは系全体の確率を条件付き確率の積で表し，条件付き確率を親節点から子節点への有向枝に対応させグラフ化している．正確にはBNはcycleが無い有向グラフ (DAG: Directed Acyclic Graph) で定義され，知識処理の分野で広く用いられている．また，統計学の分野でも，多変量解析で求められた確率分布の確率変数間の関連を，グラフで表現するグラフィカルモデリングの研究が盛んに行われている．

符号理論の場合も情報系列 $\mathbf{U}$, パリティ$\mathbf{X}$, 受信系列 $\mathbf{Y}$ の各シンボルを確率変数としてとらえ全体の確率構造を考えると，各確率変数は部分的に関連をもつ構造となっている．符号のグラフ表現としては，白丸の節点で確率変数を表し，黒丸の節点で関連の有ることを表すTannerグラフ [11] が従来から提案されていた．Tannerグラフは近年factorグラフとして一般化されている [11]．また，畳み込み符号や一部の線形符号は，トレリス [19] によってその構造を簡素に表現可能であり，従来から広く用いられている．

このようなグラフ表現は，部分的に確率変数が関連を持つ分布の確率構造を視覚的に分かりやすくする為だけではなく，このグラフを用いて事後確率計算等を効率良く行うアルゴリズムを構成するために役立っている．例えば，符号理論で用いられるトレリスは，最尤復号を保証する効率よいアルゴリズムであるViterbiアルゴリズム [19] の計算プロセスを表していると考えることもできる．後で詳しく述べるBNの確信度更新アルゴリズムとTurbo復号の同等性が明らかになった最近では [11][14]，様々な符号がBNやfactorグラフを用いて表現され，符号の性能や効率的復号法などの研究がなされている．図1はBNを用いて符号を表現した一例である．

4 事後確率の効率的計算アルゴリズム

隠れマルコフモデルにおいて，観測値が与えられたもとでの各時点の状態の事後確率を求める効率的アルゴリズムとしてforward-backwardアルゴリズム [2] がある．マルコフモデルの条件付き独立性をうまく利用し，時点順に計算する値と最終時点から時点を遡りながら計算する値を合わせるだけで事後確率を効率良く計算するアルゴリズムである．グラフ的な視点からアルゴリズムを見ると，状態遷移を表したトレリス上を順方向に流れる情報と逆方向に流れる情報が，それぞれの端点まで到達するとアルゴリズムは終了し，正しい事後確率を出力することになる．このアルゴリズムは隠れマルコフモデルが用いられる様々な分野で使用されており，応用分野の一例としはて音声認識の問題がある．

BNにおける確信度更新，つまり事後周辺確率の計算における効率的アルゴリズムはとしてはBPが代表的である．BPでは有向グラフの順方向へメッセージπと逆方向へのメッセージλを伝搬させることにより各節点の事後周辺確率を求めている．BPも事後周辺確率を部分的な確率計算の反復で行う効率的アルゴリズムと解釈することができる．BNはDAGとして定義されているが，BPを用いることができるグラフはDAGの部分クラスであるPolytree(一般木) と呼ばれるloopの無いDAGのクラスに限定されている [15]．BPが正しく事後周辺確率を計算できる保証はこの木構造上で条件付き独立の性質をうまく用いた点にあり，先に述べたforward-backwardアルゴリズムとアルゴリズムの本質は同じである．

BPの性能が保証できるクラスはこのPolytreeに限られるが，計算量は確率変数の数とアルファベットの大きさの多項式オーダとなる効率の良いアルゴリズムとなっている．Polytreeより広いクラスのDAGに対しても正確さの保証のある計算アルゴリズムが幾つか提案されているが計算量は増加してしまう [15] [13]．この問題を計算量から考えると，DAGの一般的クラスにおける事後周辺確率計算問題はNP困難となることが計算理論分野で有名な3-SAT問題を用いて証明されている [4]．

その他のグラフ表現としてfactorグラフにおいては効率的な事後周辺確率計算アルゴリズムとしてsum-productsアルゴリズムがある．グラフ上で2種類の情報を2種類の節点を通して流し計算をするアルゴリズムで，節点に集まってきた情報を和と積で計算することでこのように呼ばれる．BPと本質的には同じアルゴリズムであることが示されており，このアルゴリズムもfac-

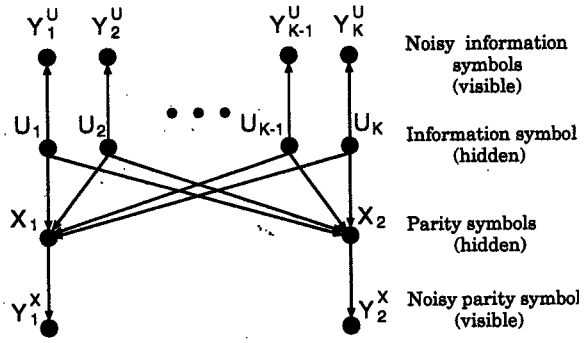

図 1: BN による符号の表現例

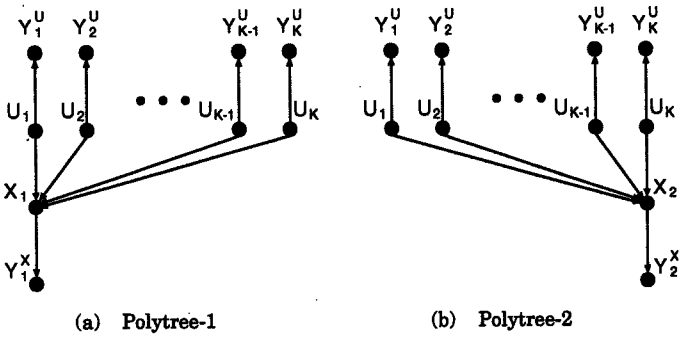

図 2: 図 1 の BN の部分グラフ

tor グラフに loop がない場合のみ正しい計算の保証がある．統計力学の分野では MRF 上での確率変数の事後期待値を効率よく計算する手法に平均場近似がある．平均場近似には様々な手法があるが，ある種の手法は BP や sum-products アルゴリズムと同等であることが知られている [12]．また，DAG よりさらに広い分布クラスに対して，与えられる情報も確率変数の値としてだけでなく確率として与えられる場合にも周辺事後確率計算が行えるアルゴリズムも提案されている [17]．

5 符号のグラフ表現と効率的復号アルゴリズム

符号のグラフ表現とその上での効率的な事後周辺確率計算アルゴリズムを眺めてみよう．例えば畳み込み符号を BN で表現すると Polytree で表現可能なため，BP アルゴリズムで正確な事後周辺確率が計算できることになる．トレリス符号に対し正確な MAP 復号を実現するアルゴリズムとして符号理論の分野では BCJR[1] アルゴリズムが有名であるが，これは forward-backward アルゴリズムそのものであることが知られている．畳み込み符号の MAP 復号は確率的には隠れマルコフモデルの状態推定と同等な問題であるので，2 つのアルゴリズムが一致することは当然の帰結といえる．さらに semi-ring 上の演算の定義を変えれば forward-backward アルゴリズムは先程述べた Viterbi アルゴリズムとも等価となる [7] [11]．これらのアルゴリズムはすべて BP の特殊形と解釈されるので，符号を BN で表現した場合に Polytree で表現可能であれば，その符号に対して BP と等価な効率よい復号アルゴリズムが構成できることになる．

しかし，オリジナルの Turbo 符号 (並列連接畳み込み符号) をはじめ多くの符号は DAG による表現は可能であるが Polytree で表現されないことが多い．例えば，図 1 の符号は loop があり Polytree とはなっていない．このような場合でも，BP をうまく用いることで，情報シンボルの事後確率の近似計算を行えないだろうか．図 1 をよく見ると，節点 X_1 に集まってくる枝に着目した部分グラフは図 2(a) のように Polytree になっていることに気づく．この部分グラフを Polytree-1 と呼び，節点 X_2 に着目した図 2(b) の部分グラフを Polytree-2 と呼ぶ．まず Polytree-1 上で BP により各節点の暫定的な事後確率計算を行う．Polytree-1 上で得られた情報系列 $\mathbf{U}$ に関する暫定的な事後確率を用いて Polytree-2 上でまた BP を行い，それを反復することで暫定事後確率の値を更新し収束させることを考える．実は BP をこのように用いた反復アルゴリズムが，Turbo 復号で行われている近似事後確率計算と同等であることが示される [14]．

オリジナルの Turbo 符号は，情報系列を一つの畳み込み符号で符号化した符号を送信し，それに並列して，その情報系列にインタリーバをかけ置換した新たな系列を，もう一つの畳み込み符号で符号化したパリティ部も送信するもので，並列連接畳み込み符号と考えられる．Turbo 符号全体の DAG を 2 つの畳み込み符号の DAG に分割すれば，それぞれは Polytree となり，BP と等価な BCJR アルゴリズムを交互に用い反復することにより，事後周辺確率の近似値が求まることになる．

また，Gallager により提案され最近再発見された LDPC (低密度パリティ検査) 符号 [5] [16] は，2 元符号で検査行列 H の各列の 1 の数を J 個と一定数に制限した符号である．J は 3 など比較的小さい数が用いられ，検査行列の 1 の密度が低いためこのように呼ばれる．この符号も DN や factor グラフで表現すると loop があるグラフとなってしまう．そこで，Turbo 復号と同様に符号を表現したグラフを幾つかの loop のない部分グラフに分解し BP または sum-products アルゴリズムで反復計算し復号を行っている．sum-products アルゴリズムを用いる場合，節点に集まってくる情報の数が J となることで和の計算の次元が J となり，効率的計算が可能となっている．統計力学のイジングスピンモデルにおける TAP

平均場近似は上記のアルゴリズムと等価であることが知られており，統計力学の分野からLDPC符号を考察した研究も行われている[12].

このように一般のDAGにBPを無理やりに用いた場合の性能は保証されるのであろうか．残念ながら，一般的には正しい事後確率を計算することの保証どころか，収束性も保証されていない．そのためTurbo復号やLDPC符号におけるsum-productsアルゴリズムは，統計学や確率推論の研究者からは性能補償範囲を超えた強引な適用法として，必ずしも良い評価ばかりではない．しかし，これらの2つの符号をはじめいくつかの符号に対しては，復号誤り特性に関する多くのシミュレーションの結果から，これらの復号法の良好な性能が確認されている．理論的面からは，BPの演算を変換したBPと同質なアルゴリズム[3]については、いくつかの性質が保証されている。例えば、loopが一つでアルファベットが2元のDAGに対しアルゴリズムは収束し，収束結果は真の最大事後確率による推定と一致するという意味で正当性が示されている[20]．この結果を拡張して単一loopが複数存在するDAGに対しても，ある範囲内で正当性の保証が得られることが最近報告されている．また最近，情報幾何を用いてTurbo復号の収束性について考察する研究[9]も行われている．

6 まとめ

事後確率，事後期待値を求める問題は，知識処理，符号理論，統計学，情報理論，学習理論，統計力学，制御理論など様々な分野で重要な問題であり，グラフによる確率分布の表現法や計算アルゴリズムも多岐にわたっている．本稿では，BPを用いた符号理論における復号問題を中心にこれらの研究の関連についてほんの一部を垣間見た．今後，これらの各分野が益々相互に関連をもち，この問題に対しての研究がさらに盛んになることが期待されている．

参考文献

[1] L.R. Bahl, J. Cocke, F. Jelinek and J. Raviv, *Optimal decoding of linear codes for minimizing symbol error rate* IEEE Trans. IT, Vol.20, 1974.

[2] L.E. Baum and T. Petrie, *Statistical inference for probabilistic functions on finite state markov chains* Annals of Mathematical Statistics, Vol.37, 1966.

[3] C. Berrou, A. Glavieux and P. Thitimajshima, *Near Shannon limit error-correcting coding and decoding*, in Proc. IEEE Int. Conf. Commun., 1993.

[4] P. Dagum, *Approximating Probabilistic Inference in Bayesian belief networks is NP-hard* Artificial Intelligence, Vol.42, 1990.

[5] R.G. Gallager, *Low-Density Parity-Check Codes* MIT Press, Cambridge, 1963.

[6] J. Hagenauer, E. Offer and L. Papke, *Iterative decoding of binary block and convolutional codes*, IEEE Trans. IT, Vol.42, 1996.

[7] C. Heegard and S.B. Wicker, *Turbo coding* Kluwer Academic Publishers, 1999.

[8] 平澤茂一，西島利尚, 符号理論入門, 培風館, .

[9] 池田思朗，田中利幸, 甘利俊一, ターボ符号の情報幾何, IBIS2001, 2001.

[10] 井坂元彦, 今井秀樹, *Shannon*限界への道標: *"Parallel concatenated (Turbo) coding", "Turbo (Iterative) decoding"*とその周辺, 信学技報IT98-51, 1998.

[11] F.R. Kschischang and B.J. Frey, *Iterative decoding of compound codes by probability propagation in graphical models*, IEEE J. Sel. Areas Commun., Vol.16 No.2, 1998.

[12] Y. Kabashima and D. Sad, *Belief propagation vs. TAP for decoding corrupted messages*, Europhys. Lett. Vol.44, No.5, 1998.

[13] S.L. Lauritzen and D.J. Spiegelhalter, *Local computation with probabilities on graphical structures and their application to expert systems*, J.R. Statist. Soc., Vol.50, No.2, 1988.

[14] R.J. McElice, D.J.C. MacKay and J. Cheng, *Turbo decoding as an instance of Pearl's "Belief Propagation"*, IEEE J. Sel. Areas Commun., Vol.16, No.2, 1998.

[15] J. Pearl, *Probabilistic reasoning in intelligent systems* Morgan Kaufmann, 1988.

[16] D.J.C. MacKay, *Good Error-Correcting Codes Based on Very Sparse Matrices* , IEEE IT., Vol.45, No.2, 1999.

[3]本稿で中心的に述べた事後周辺確率を求めるのではなく、未知確率変数ベクトルの値を事後確率最大で推定することが目的のアルゴリズム

[17] T. Matsushima, T.K. Matsushima and S. Hirasawa, *An Iterative Calculation Algorithm for Posterior Probability*, Proc. the 23rd Symp. on Information Theory and Its Applications, 2000.

[18] 本村陽一, ベイジアンネットワーク, 電子情報通信学会誌, Vol.83, No.8, 2000

[19] A. Viterbi and J.K. Omura, *Principle of Digital Communication and Coding*, McGraw-Hill, New York, 1979.

[20] Y. Weiss, *Belief propagation and revision in networks with loop*, M.I.T A.I. Memo No.1616, 1997.

The 24th Symposium on Information Theory
and Its Applications (SITA2001)
Kobe, Hyogo, Japan, Dec. 4 - 7, 2001

ブロックターボ符号に対するインタリーバの構成法と最小距離
On the Construction of an Interleaver for Block Turbo Codes and Minimum Distance

小林 学* 松嶋 敏泰† 平澤 茂一†
Manabu KOBAYASHI Toshiyasu MATSUSHIMA Shigeichi HIRASAWA

Abstract— In 1993 C. Berrou et. al. have proposed turbo codes which achieved low BER with a SNR per information bit close to Shannon's theoretical limit on AWGN channel. H. Hagenauer et. al. also have proposed block turbo codes which consist of two systematic block codes concatenated in parallel. In this paper we show the lower bound of minimum distance by restricting an interleaver. Furthermore we propose an algorithm to construct the component codes and an interleaver which maximize its lower bound. Finary we show that the minimum distance of block turbo codes increases in comparison with conventional one.

Keywords—Block turbo codes, Minimum distance, Generator matrix, Interleaver

1 まえがき

1993 年 C.Berrou らは，AWGN 通信路に対し単位情報記号あたりの信号対雑音比 (E_b/N_0) に対する Shannon 限界に近いビット誤り確率 (BER) を達成するターボ符号を提案した．C.Berrou らにより提案されたターボ符号は帰還回路を持つ 2 つの組織畳込み符号器を並列に連接した符号である [1]．さらに要素符号にブロック符号を用いたブロックターボ符号も提案されている [2, 4]．

本稿では H.Hagenauer らの提案したブロックターボ符号 [2] に対しインタリーバに制限を加えることにより，情報記号の重みが 1 および 2 のときのブロックターボ符号の符号語に対する Hamming 重みの下界を求める．さらにこの下界を大きくするインタリーバの構成手法を示し，ブロックターボ符号の最小距離を大きくすることが可能な要素符号を探索により求めるアルゴリズムを提案する．結果的に得られる要素符号およびインタリーバを用いることにより，大きな最小距離を保証するブロックターボ符号を設計することが可能となる．

2 Hagenauer 型ブロックターボ符号

H.Hagenauer らにより提案されたブロックターボ符号 C_H は，2 つの要素符号の検査記号が情報記号にのみ依存し，互いの検査記号には影響を受けない構成となっている．ここでは (n_1, k_1, d_1) 組織ブロック符号 C_1 と (n_2, k_2, d_2) 組織ブロック符号 C_2 を要素符号としたときのブロックターボ符号の符号化について述べる．まず k_1k_2 シンボルの情報記号 $\boldsymbol{u}$ を k_1 シンボル毎に k_2 個のベクトルに分割する．すなわち $\boldsymbol{u} = (\boldsymbol{u}_1^{(1)}, \boldsymbol{u}_2^{(1)}, \cdots, \boldsymbol{u}_{k_2}^{(1)})$, $\boldsymbol{u}_i^{(1)} = (u_{i,1}^{(1)}, u_{i,2}^{(1)}, \cdots, u_{i,k_1}^{(1)}), i = 1, 2, \ldots, k_2$, と置く．それぞれのベクトル $\boldsymbol{u}_i^{(1)}$ を情報記号とみなし，組織符号を用いて符号化を行う．本稿では簡単のため，それぞれを同一の C_1 を用いて符号化を行うものと仮定する[1]．また C_1 の生成行列を G_1 と表すとこの符号化により $\boldsymbol{c}^{(1)} = (\boldsymbol{c}_1^{(1)}, \boldsymbol{c}_2^{(1)}, \ldots, \boldsymbol{c}_{k_2}^{(1)})$, $\boldsymbol{c}_i^{(1)} = (\boldsymbol{u}_i^{(1)}, \boldsymbol{x}_i^{(1)})$, $\boldsymbol{x}_i^{(1)} = (x_{i,1}, x_{i,2}, \ldots, x_{i,n_1-k_1})$, が得られる．ただし $\boldsymbol{c}_i^{(1)} = \boldsymbol{u}_i^{(1)} G_1$ であり，$G_1 = [I_{k_1}, Q_1]$ の形式をしているものとする．ここで I_a は $a \times a$ 単位行列を表し，Q_1 は $k_1 \times (n_1 - k_1)$ 行列である．

さらにインタリーバにより $\boldsymbol{u}$ を置換し，今度はこれを k_2 シンボル毎に k_1 個のベクトルに分割する．これを $\boldsymbol{u}' = (\boldsymbol{u}_1^{(2)}, \boldsymbol{u}_2^{(2)}, \ldots, \boldsymbol{u}_{k_1}^{(2)})$, $\boldsymbol{u}_i^{(2)} = (u_{i,1}^{(2)}, u_{i,2}^{(2)}, \ldots, u_{i,k_2}^{(2)}), i = 1, 2, \ldots, k_1$, と表す．このそれぞれのベクトルに対し同一の C_2 を用いて符号化を行う．C_2 の生成行列を G_2 と表すと結果的に $\boldsymbol{c}^{(2)} = (\boldsymbol{c}_1^{(2)}, \boldsymbol{c}_2^{(2)}, \ldots, \boldsymbol{c}_{k_1}^{(2)})$, $\boldsymbol{c}_i^{(2)} = \boldsymbol{u}_i^{(2)} G_2$, が得られる．最終的に Hagenauer 型のブロックターボ符号の符号語 $\boldsymbol{c}_H$ は次式で表すことができる．

$$\boldsymbol{c}_H = (\boldsymbol{c}_1^{(2)}, \boldsymbol{c}_2^{(2)}, \ldots, \boldsymbol{c}_{k_1}^{(2)}, \boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \ldots, \boldsymbol{x}_{k_2}^{(1)}). \quad (1)$$

さてこの Hagenauer 型のブロックターボ符号の生成行列を示す．$\boldsymbol{u}$ から $\boldsymbol{c}^{(1)}$ を求めるためには，$k_1k_2 \times k_2n_1$ 行列 $G_H^{(1)}$ を

$$G_H^{(1)} = \begin{bmatrix} G_1 & & & \mathbf{O} \\ & G_1 & & \\ & & \ddots & \\ \mathbf{O} & & & G_1 \end{bmatrix}, \quad (2)$$

と置くと $\boldsymbol{c}^{(1)} = \boldsymbol{u} G_H^{(1)}$ となる．次にインタリーバを $k_2n_1 \times k_2n_1$ 置換行列 $P_H = [p_{a,b}]$ により表現する．置換行列とは各行，各列の Hamming 重みが 1 の行列である．ここで C_1 の検査記号の各ベクトルを後ろへ置換するために P_H を次のように制限する．

$$p_{(i-1)n_1+k_1+j, k_1k_2+(i-1)(n_1-k_1)+j} = 1, \quad \forall i \in [1, k_2], \forall j \in [1, n_1 - k_1], \quad (3)$$

ただし整数 r, s に対し $[r, s] = \{r, r+1, \ldots, s\}$ と定義する．これにより

$$\boldsymbol{u} G_H^{(1)} P_H = (\boldsymbol{u}', \boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \ldots, \boldsymbol{x}_{k_2}^{(1)}), \quad (4)$$

とすることができる．最後に式 (4) を符号化して式 (1) とするためには，$k_2n_1 \times \{k_1n_2 + (n_1 - k_1)k_2\}$ 行列 $G_H^{(2)}$ を

$$G_H^{(2)} = \begin{bmatrix} \overbrace{\begin{matrix} G_2 & & \\ & G_2 & \\ & & \ddots \\ & & G_2 \end{matrix}}^{k_1\text{個}} & & \mathbf{O} \\ \mathbf{O} & & I_{(n_1-k_1)k_2} \end{bmatrix}, \quad (5)$$

と置けばよい．結果的に Hagenauer 型のブロックターボ符号 C_H の生成行列 G_H は式 (2),(3),(5) を用いて $G_H = G_H^{(1)} P_H G_H^{(2)}$ となる．

* 早稲田大学理工総合研究所，〒169-8555 新宿区大久保 3-4-1, School of Science and Engineering, Waseda University, 3-4-1 Ohkubo Shinjyuku-ku, Tokyo, 169-8555 Japan, E-mail:manabu@hirasa.mgmt.waseda.ac.jp
† 早稲田大学理工学部経営システム工学科，同上．
[1] より一般に可変符号化に拡張することは容易である．

$P_H=[p_{a,b}]$ に式(3)以外の制約を次のように設ける．

$$\sum_{a=(i-1)n_1+1}^{(i-1)n_1+k_1}\ \sum_{b=(j-1)k_2+1}^{jk_2} p_{a,b} = 1, \qquad (6)$$
$$\forall i\in[1,k_2], \qquad \forall j\in[1,k_1],$$

ただし式(6)の和は通常の整数の加算を表す．置換行列により $u^{(1)}_{i_1,y_1}, u^{(1)}_{i_2,y_2}$ がそれぞれ $u^{(2)}_{j_1,z_1}, u^{(2)}_{j_2,z_2}$ に置換されたとすると，式(6)の制約により $i_1=i_2$ ならば $j_1\neq j_2$ となる．本稿では式(6)を満足する置換行列 P_H を持つ Hagenauer 型ブロックターボ符号 C^*_H を対象とする．

定義 1 ある符号 C に対し情報記号の Hamming 重みが w でかつ検査記号の Hamming 重みが z の符号語数を $A^C_{w,z}$ とする．また $W^C_{min}(w)=\min_z\{w+z|A^C_{w,z}>0\}$ と定義する． □

このとき次の定理が成り立つ．

定理 1 $d_1=d_2$ のとき $H(w)$ を次式で定義する．

$$H(w)=\begin{cases}(2\lceil\sqrt{w}\rceil-1)d_1-\lceil\sqrt{w}\rceil(\lceil\sqrt{w}\rceil-1), \\ \qquad\qquad \text{if } (\lceil\sqrt{w}\rceil-1)\lceil\sqrt{w}\rceil\geq w \\ 2\lceil\sqrt{w}\rceil d_1-\lceil\sqrt{w}\rceil^2, \qquad \text{otherwise}\end{cases} \qquad (7)$$

ただし $\lceil a\rceil$ は a 以上の最小の整数を表す．このとき

$$\min_{w'\geq w} W^{C^*_H}_{min}(w')\geq H(w), \qquad (8)$$

が成り立つ．

(証明) 文献[5]定理5,7参照． □

3 低重み情報に対する符号語の重みの下界

本節では情報記号の Hamming 重み $w=1,2$ それぞれに対する C^*_H の符号語の最小 Hamming 重み $W^{C^*_H}_{min}(w)$ の厳しい下界を示す．まずインタリーバ（置換行列）P_H にさらなる制限を加え，$W^{C^*_H}_{min}(w), w=1,2,$ の下界を示す．次にこの下界を大きくするインタリーバの構成法について述べる．以降簡単のため要素符号を $C_1=C_2$ とし，かつその生成行列を $G_1=G_2$ とする．

定義 2 ある正定数 x_{max} に対し $F:[1,x_{max}]\to[1,x_{max}]$ を $F^{-1}(i)=F(i), i\in[1,x_{max}]$, を満たす全単射の関数とする．生成行列 G_1 の各行を $\boldsymbol{g}_i, i=1,2,\ldots,k_1$, と表し，$X_1,X_2,\ldots,X_{x_{max}}$ を

$$\bigcup_{i=1}^{x_{max}} X_i=[1,k_1], \qquad X_i\cap X_j=\phi,\ i\neq j,$$
$$|X_j|=|X_{F(j)}|\neq 0, \qquad j=1,2,\ldots,x_{max}, \qquad (9)$$

を満たすように G_1 の行番号の集合を分割したものとする．ただし $|X|$ は集合 X の要素数を表す．さらに関数 $J(y)$ を $y\in X_i\Leftrightarrow J(y)=i$ と定義し，$P(y)$ を次式で定義する．

$$P(y)=\min_{j\in X_{F(J(y))}}\{w_H(\boldsymbol{g}_j)-1\}. \qquad (10)$$

ただし $w_H(\boldsymbol{a})$ は $\boldsymbol{a}$ の Hamming 重みとする． □

このとき置換行列 $P_H=[p_{a,b}]$ に対し式(3),(6)に加え，さらに次のような制約を設ける．

$$(a \bmod n_1)\in X_i \text{ and } p_{a,b}=1$$
$$\Rightarrow (b-1 \bmod k_2)+1\in X_{F(i)}. \qquad (11)$$

すなわち，C_1 における X_i の要素の情報記号は C_2 における $X_{F(i)}$ の要素の情報記号へと置換する．このとき $W^{C^*_H}_{min}(1)$ に関して次の補題が成り立つ．

補題 1 式(3),(6),(11)を満足するように置換行列に制限を加えた時，次式を満足する定数 D_1 が存在するならば $W^{C^*_H}_{min}(1)\geq D_1$ が成り立つ．

$$P(y)+w_H(\boldsymbol{g}_y)\geq D_1,\ \forall y\in[1,k_1]. \qquad (12)$$

(証明) $w_H(\boldsymbol{u})=1$ を満たす $\boldsymbol{u}$ について考える．ある i,y に対し $\boldsymbol{u}^{(1)}_i=(u^{(1)}_{i,1},u^{(1)}_{i,2},\ldots,u^{(1)}_{i,k_1})$, $u^{(1)}_{i,y}=1$ とすると $G^{(1)}_H$ による符号化により $\boldsymbol{c}^{(1)}_i=\boldsymbol{g}_y$ となる．また $a=(i-1)n_1+y$ に対し $p_{a,b}=1$ とすると，$u^{(1)}_{i,y}$ は置換行列により $u^{(2)}_{j,z}$, $j=\lceil b/k_2\rceil$, $z=(b-1 \bmod k_2)+1$ へ置換される．従って $u^{(2)}_{j,z}=1$ より $\boldsymbol{c}^{(2)}_j=\boldsymbol{g}_z$ となる．ここで式(11)より $z\in X_{F(J(y))}$ であるから，式(10)の定義より $\boldsymbol{c}^{(2)}_j$ の検査記号の Hamming 重みについて $w_H(\boldsymbol{g}_z)-1\geq P(y)$ が成り立つ．従って

$$w_H(\boldsymbol{c}_H)=w_H(\boldsymbol{c}^{(1)}_i)+w_H(\boldsymbol{c}^{(2)}_j)-1\geq w_H(\boldsymbol{g}_y)+P(y), \qquad (13)$$

である．$u^{(1)}_{i,y}=1$ の i,y は任意として式(13)は成り立つので，補題が成り立つ． □

次にある正定数 D_1 に対し式(12)を満足するような集合 $X_1,X_2,\ldots,X_{x_{max}}$ を求める．そのために $Q(y):=\min_j\{j\geq D_1-w_H(\boldsymbol{g}_y)|A^{C_1}_{1,j}>0\}, y\in[1,k_1]$, $S_z:=\{y\in[1,k_1]|\ w_H(\boldsymbol{g}_y)-1=z\}, z\in[d_1-1,n_1-k_1]$, $x_{max}:=0$ とし，次のアルゴリズムを実行する．ただし $\lfloor a\rfloor$ は a 以下の最大の整数とする．

[探索アルゴリズム1]

(1) $z=d_1-1,d_1,\ldots,\lfloor\frac{D_1-2}{2}\rfloor$ について以下を行う．

(i) $S_z\neq\phi$ ならば以下を行う．

(ii) もし $q_i=z, \exists i\in[1,x_{max}]$ ならば $m:=i$, $flag:=0$ とする．そうでなければ $x_{max}:=x_{max}+2$, $F(x_{max}-1)=x_{max}, F(x_{max})=x_{max}-1, X_{x_{max}-1}=X_{x_{max}}:=\phi$, $m:=x_{max}-1$, $flag:=1$ とする．

(iii) 適当な $y\in S_z$ 1個に対し $X_m:=X_m\cup\{y\}$, $S_z:=S_z\setminus\{y\}$. また $t\geq Q(y)$ に対し $l\in S_t$ とし $X_{F(m)}:=X_{F(m)}\cup\{l\}$, $S_t:=S_t\setminus\{l\}$. もし $flag=1$ ならば $q_{x_{max}-1}:=z$, $q_{x_{max}}:=Q(y)$ とする．(i)へ．

(2) $Z:=\left\{z\geq\lceil\frac{D_1-1}{2}\rceil\,|S_z\neq\phi\right\}$ とし，もし $Z\neq\phi$ ならば $x_{max}:=x_{max}+1$, $X_{x_{max}}:=\bigcup_{z\in Z}S_z$, $F(x_{max}):=x_{max}$ とする．$X_1,X_2,\ldots,X_{x_{max}}$ を出力して終了． □

定理 2 正定数 D_1 に対して

$$\sum_{i=d_1-1}^{z}A^{C_1}_{1,i}\leq\sum_{i=D_1-z-1}^{n_2-k_2}A^{C_2}_{1,i},\ \forall z\in[d_1-1,n_1-k_1], \qquad (14)$$

が成り立つならば式(12)を満足する $X_1,X_2,\ldots,X_{x_{max}}$ が存在し，探索アルゴリズム1により得られる．

(証明) 探索アルゴリズム1のステップ(iii)における z に対し，式(14)を満足するならば $y\in S_z$ に対し $S_t\neq\phi$ を満たす $t\geq Q(y)$ が必ず存在する．従って式(14)が成り立つときアルゴリズム1は正常に終了する．次に，ステップ(iii)において $y\in X_m$ に対し $w_H(\boldsymbol{g}_y)-1=z$ であり，かつ $l\in X_{F(m)}$ に対し $w_H(\boldsymbol{g}_l)-1\geq Q(y)$ であるから $P(y)\geq Q(y)$ が成り立つ．また $w_H(\boldsymbol{g}_l)-1\geq Q(y)\geq D_1-w_H(\boldsymbol{g}_y)=D_1-(z+1)$ より $P(l)=z\geq D_1-w_H(\boldsymbol{g}_l)$ も成り立つ．

さらにステップ (2) において $z \geq \lceil \frac{D_1-1}{2} \rceil, z \in Z$, であるから，$y \in X_{x_{max}}$ に対し $P(y) \geq \lceil \frac{D_1-1}{2} \rceil$ である．従って $w_H(\boldsymbol{g}_y) + P(y) \geq 2\lceil \frac{D_1-1}{2} \rceil + 1 \geq D_1$ が成り立つ．以上より定理が成り立つ． □

次に $W_{min}^{C_H^*}(2)$ の下界に関して次の補題が成り立つ．

補題 2 式(3),(6),(11)を満足するように置換行列に制限を加えた時，次式を満足する定数 $D_2, 2D_1 \geq D_2$, が存在するならば $W_{min}^{C_H^*}(2) \geq D_2$ が成り立つ．

$$P(y) + P(z) + w_H(\boldsymbol{g}_y + \boldsymbol{g}_z) \geq D_2, \quad \forall y, z \in [1, k_1], y \neq z. \tag{15}$$

(証明) $w_H(\boldsymbol{u}) = 2$ を満たす $\boldsymbol{u}$ について考える．$i_l, y_l, l = 1, 2,$ に対し $u_{i_l,y_l}^{(1)} = 1$ とし，また $a_l = (i_l - 1)n_1 + y_l$ について $p_{a_l,b_l} = 1$ とする．このとき補題1の証明と同様 $u_{i_l,y_l}^{(1)}$ は置換行列により $u_{j_l,z_l}^{(2)}$, $j_l = \lceil b_l/k_2 \rceil$, $z_l = (b_l - 1 \bmod k_2) + 1$ へ置換される．まず $i_1 = i_2$ を仮定すると $\boldsymbol{c}_{i_1}^{(1)} = \boldsymbol{g}_{y_1} + \boldsymbol{g}_{y_2}$ となる．また式(6)の制限より $j_1 \neq j_2$ であるから

$$\begin{aligned} w_H(\boldsymbol{c}_H) &= w_H(\boldsymbol{c}_{i_1}^{(1)}) + w_H(\boldsymbol{c}_{j_1}^{(2)}) - 1 + w_H(\boldsymbol{c}_{j_2}^{(2)}) - 1 \\ &\geq w_H(\boldsymbol{g}_{y_1} + \boldsymbol{g}_{y_2}) + P(y_1) + P(y_2), \end{aligned} \tag{16}$$

が成り立つ．次に $i_1 \neq i_2$ を仮定する．もし $j_1 \neq j_2$ ならば

$$\begin{aligned} w_H(\boldsymbol{c}_H) = {} & w_H(\boldsymbol{c}_{i_1}^{(1)}) + w_H(\boldsymbol{c}_{i_2}^{(1)}) + w_H(\boldsymbol{c}_{j_1}^{(2)}) - 1 \\ & + w_H(\boldsymbol{c}_{j_2}^{(2)}) - 1 \\ \geq {} & w_H(\boldsymbol{g}_{y_1}) + w_H(\boldsymbol{g}_{y_2}) + P(y_1) + P(y_2), \end{aligned} \tag{17}$$

が成り立つ．ここで $y_1 \neq y_2$ ならば

$$\text{式(17)の最右辺} \geq w_H(\boldsymbol{g}_{y_1} + \boldsymbol{g}_{y_2}) + P(y_1) + P(y_2), \tag{18}$$

であり，$y_1 = y_2$ ならば

$$\text{式(17)の最右辺} \geq 2D_1, \tag{19}$$

となる．また $j_1 = j_2$ ならば

$$\begin{aligned} w_H(\boldsymbol{c}_H) &= w_H(\boldsymbol{c}_{i_1}^{(1)}) + w_H(\boldsymbol{c}_{i_2}^{(1)}) + w_H(\boldsymbol{c}_{j_1}^{(2)}) - 2 \\ &\geq P(z_1) + P(z_2) + w_H(\boldsymbol{g}_{z_1} + \boldsymbol{g}_{z_2}), \end{aligned} \tag{20}$$

が成り立つ．式(20)の最後の不等式は関数 $F = F^{-1}$ の対称性からなる．以上より，補題が成り立つ． □

次に符号 $C_1 (= C_2)$ が式(14)を満足しているとき，式(15)を満足するような集合 $X_1, X_2, \ldots, X_{x_{max}}$ を探索により求めるアルゴリズムを以下に示す．

[探索アルゴリズム 2]

(初期化)

(1) $y = 1, 2, \ldots, k_1$ に対し $Q(y) := \min_j \{j \geq D_1 - w_H(\boldsymbol{g}_y) | A_{1,j}^{C_1} > 0\}$.

(2) $z = d_1 - 1, d_1, \ldots, n_1 - k_1$ に対し $S_z := \{y \in [1, k_1] | w_H(\boldsymbol{g}_y) - 1 = z\}$, $\Delta_z := \sum_{i=D_1-z-1}^{n_1-k_1} A_{1,i}^{C_1} - \sum_{i=d_1-1}^{z} A_{1,i}^{C_1}$.

(3) $M := \{\{y, z\} | Q(y) + Q(z) + w_H(\boldsymbol{g}_y + \boldsymbol{g}_z) < D_2, 1 \leq y < z \leq k_1\}$, $E := \phi$, $x_{max} := 0$.

(探索)

(4) もし $M = \phi$ ならば (12) へ．

(5) $y = 1, 2, \ldots, k_1$ に対し $M_y := \{\{y, z\} \in M\}$.

(6) $L := \{y \in [1, k_1] | M_y \neq \phi\}$.

(7) もし任意の $y, z \in E$, に対し，ある $\{y, z\} \in M$ が存在するならば探索失敗として終了．

(8) $l := \arg\max_{y \in L \setminus (E \cap L)} |M_y|$, $t_0 := \max_{y | \{l,y\} \in M_l} \{D_2 - Q(y) - w_H(\boldsymbol{g}_y + \boldsymbol{g}_l)\}$, $t_1 := \min\{y \geq t_0 | S_y \neq \phi\}$.

(9) もし $i = Q(l), Q(l)+1, \ldots, t_1 - 1$ 全てに対し $\Delta_i > 0$ ならば $i = Q(l), Q(l) + 1, \ldots, t_1 - 1$ それぞれについて $\Delta_i := \Delta_i - 1$. $Q(l) := t_1$, $M := M \setminus M_l$ とし，(10)へ．そうでなければ $E := E \cup \{l\}$ とし，(7)へ．

(10) もし $q_i = Q(l), \exists i \in [1, x_{max}]$ ならば $m := i$, $flag := 0$ とする．そうでなければ $x_{max} := x_{max} + 2$, $F(x_{max} - 1) = x_{max}$, $F(x_{max}) = x_{max} - 1$, $X_{x_{max}-1} = X_{x_{max}} := \phi$, $m := x_{max} - 1$, $flag := 1$ とする．

(11) 適当な $y \in S_{Q(l)} \setminus (S_{Q(l)} \cap L)$ 1個に対し $X_m := X_m \cup \{y\}$, $S_{Q(l)} := S_{Q(l)} \setminus \{y\}$. $X_{F(m)} := X_{F(m)} \cup \{l\}$, $S_{w_H(\boldsymbol{g}_l)-1} := S_{w_H(\boldsymbol{g}_l)-1} \setminus \{l\}$. もし $flag = 1$ ならば $q_{x_{max}-1} := Q(l)$, $q_{x_{max}} := Q(y)$ とする．(4)へ．

(12) 探索アルゴリズム 1を行い終了． □

この探索アルゴリズム 2 に対し次の定理が成り立つ．

定理 3 探索アルゴリズム 2 により得られた $X_1, X_2, \ldots, X_{x_{max}}$ は式(12),(15)を満足する．

(証明) 探索アルゴリズム2において，$\forall (y, z) \in M$ に対し $P(y) = Q(y)$, $P(z) = Q(z)$ とすると式(15)を満足しない．そこでステップ (8) から (11) において $P(l) = t_1$ を満足するように X_m および $X_{F(m)}$ を生成している．これにより $\forall (l, y) \in M_l$ に対し $P(l) + P(y) + w_H(\boldsymbol{g}_l + \boldsymbol{g}_y) \geq D_2$ が成り立つため，$M := M \setminus M_l$ として繰り返している．また明らかにこの $P(l)$ は式(12)を満足している．以上よりステップ (12) が行われる時は式(15)を満足している．ここでステップ (12) においては定理 2 の証明と同様式(12)を満足するように X_m および $X_{F(m)}$ が作成される．しかしステップ (12) を行う場合式(14)と同様に

$$\begin{aligned} \Delta_z = {} & \sum_{i=D_1-z-1}^{n_1-k_1} |S_i| - \sum_{i=d_1-1}^{z} |S_i| \geq 0, \\ & \forall z \in [d_1 - 1, n_1 - k_1], \end{aligned} \tag{21}$$

を満足している必要がある．そこでステップ (9) において l に対し $P(l) = t_1$ としたとき式(21)が成り立つかどうかをチェックしている．もし式(21)が成り立たないなら $P(l) \leq t_1$ とせざるを得ないため $E := E \cup \{l\}$ とする．従ってステップ (7) において $y, z \in E$ に対し $\exists (y, z) \in M$ ならば $P(y) + P(z) + w_H(\boldsymbol{g}_y + \boldsymbol{g}_z) < D_2$ となってしまうため，アルゴリズム失敗として終了となる．以上よりアルゴリズムが正常に終了した場合，生成された $X_1, X_2, \ldots, X_{x_{max}}$ に対し式(12),(15)が成り立つ． □

以上より，与えられた符号 C_1 と生成行列 G_1 に対し，式(14)を満たす最大の D_1 を求め，また D_2 を適当な値から始めて探索アルゴリズム 2 を適用し，$X_1, X_2, \ldots, X_{x_{max}}$ が得られるたびに D_2 を上昇させることを繰り返し，探索アルゴリズム 2 が失敗するまで続ける．このとき定理 2, 3 より低重み情報記号に対する符号語の Hamming 重みの下界を大きくするインタリーバの構成が得られる．

4 最小距離の最大化

本節では C_H^* を構成したとき，最小距離が大きくなる要素符号 $C_1 (= C_2)$ を見つけることを考える．そのために任意のブロック符号 C の生成行列を列置換した符号を $R_{\max}$ 回生成し，定理 2 および探索アルゴリズム 2 を用

表 1: D_1 および D_2 を最大とする $(n_1 = 2^m - 1, k_1, d_1 = 5)$ BCH 符号の生成行列 G_1^* の例

(n_1, k_1)	D_1	D_2	Q_1^* の各行 $\boldsymbol{q}_i^*, i \in [1, k_1]$, (8進数表示)
(15,7)	10	14	127,074,057,352,266,225,162
(31,21)	13	14	0773,1535,1766,0075,1315,1716,1267,1770,1443,1557,1341,1623,0524,1405,0567,1454,0635, 1651,1106,0607,1037
(63,51)	15	16	2151,6165,0275,6267,1721,7373,4273,3407,5511,3625,1633,3715,7471,3544,2745,1364,6462, 7703,7345,7366,1263,1461,5354,3162,7041,6143,7734,5457,7533,5516,5574,3577,6777,5771, 7032,0770,4530,6527,5732,3327,2626,3752,7550,3466,6721,1072,0766,1757,7124,1335,2137
(127,113)	16	16	03650,03154,32370,26077,07071,07673,03227,14752,07563,21166,05316,15636,11647,12361, 24570,27605,05037,30474,23273,10633,23744,16105,26774,26355,03766,33141,01142,00374, 27416,02350,32057,37153,22356,33646,27211,26635,33634,23330,05577,02037,06236,05622, 21703,16534,05531,17056,23561,36433,31052,35751,02772,23427,32237,37454,33733,00453, 12630,36614,16741,13176,06366,36273,00216,01326,07160,31276,27727,17217,27103,25333, 12112,32672,37323,10455,25655,26034,14661,24314,20652,27336,35245,13345,34070,37354, 33431,32547,34236,17760,20223,27460,24720,05054,13006,36352,36224,31125,24676,06223, 20071,27447,31615,14315,05502,27346,35660,37477,17704,12677,13252,35765,06713,16337, 30706

いることにより，補題 1 , 2 を満足する最も大きい D_1 および D_2 を持つ C_H^* の構成を求めるアルゴリズムを以下に示す．

[生成行列生成アルゴリズム]

(1) (n_1, k_1, d_1) 符号 C に対する生成行列 G を求め，$D_1^* := 0$, $D_2^* := 0$, $repeat := 0$ とする．

(2) $repeat = R_{\max}$ ならば G_1^*, $X_1^*, X_2^*, \ldots, X_{x_{max}^*}^*$, D_1^*, D_2^* を出力して終了．

(3) G の列をランダムに置換した行列を作成し，その行列のはじめの k 列を線形独立となるように列置換する．さらに，この行列に対しはじめの k_1 列が単位行列となるように行基本操作を施し，これを G_1 とする．またこの符号を C_1 と表す．

(4) G_1 より $A_{1,i}^{C_1}$ を求め，定理 2 を用いて最大の D_1 を求める．$D_1 \geq D_1^*$ ならば (5) へ．そうでなければ $repeat := repeat + 1$ として (2) へ．

(5) 探索アルゴリズム 2 を繰り返し用いて最大の D_2 とそのときの $X_1, X_2, \ldots, X_{x_{max}}$ を求める．

(6) $D_2 > D_2^*$ ならば $G_1^* := G_1$, $x_{max}^* := x_{max}$, $X_i^* := X_i, i \in [1, x_{max}]$, $D_1^* := D_1$, $D_2^* := D_2$ とする．$repeat := repeat + 1$ として (2) へ． □

上のアルゴリズムで C に $n_1 = 2^m - 1, d_1 = 5$ を満たす (n_1, k_1, d_1) 原始 BCH 符号を用いた時の結果を表 1に示す．アルゴリズムで得られた生成行列を $G_1^* = [I_{k_1}, Q_1^*]$ と表し，さらに Q_1^* の各行を $\boldsymbol{q}_i^* \in \{0,1\}^{n_1 - k_1}, i \in [1, k_1]$, と表す．表 1にはこの $\boldsymbol{q}_i^* \in \{0,1\}^{n_1 - k_1}$ を下位から 3 ビット毎に区切り，8 進数として表示している．例えば表 1の (15,7) 符号における $\boldsymbol{q}_1^* = (0,1,0,1,0,1,1,1)$ は 127 と表記する．

表から，符号長 n_1 が増加するに従い D_1 および D_2 を大きくすることが可能であることが分かる．これは d_1 一定のもとで符号長が増加すると $n_1 - k_1$ も増加するため，$A_{1,i}^{C_1} > 0$ なる i が $d_1 - 1$ から $n_1 - k_1$ の間に広く分布する傾向があるからである．

ここでブロックターボ符号 C_H^* の最小距離に関する定理を以下に示す．

定理 4 式 (3),(6),(11) を満足するインタリーバ（置換行列）を用いて構成した C_H^* の最小距離 D_H^* について

$$D_H^* \geq \min\{D_1, D_2, H(3)\}, \tag{22}$$

が成り立つ．

(証明) 補題 1 より $W_{min}^{C_H^*}(1) \geq D_1$, 補題 2 より $W_{min}^{C_H^*}(2) \geq D_2$, 定理 1 より $w \geq 3$ に対し $W_{min}^{C_H^*}(w) \geq H(3)$ が成り立つ．明らかに $D_H^* = \min_w W_{min}^{C_H^*}(w)$ であるから，定理が成り立つ． □

$d_1 = 5$ の場合 $H(3) = 16$ であるから，表 1の符号を要素符号としたブロックターボ符号 C_H^* では $D_H^* \geq D_1$ となる．また定理 1 から明らかな通り，$H(3) = 4d_1 - 4$ は d_1 が増加するに従い線形に増加する．このとき定理 4 より最小距離は D_1 あるいは D_2 に大きく依存する．従って本節で述べた生成行列生成アルゴリズムは C_H^* を設計する上で重要な役割を演ずる．

従来の置換行列に式 (6),(11) の制約を課さないインタリーバを用いた場合，ブロックターボ符号 C_H の最小距離の下界は $d_1 (= d_2)$ である [5]．さらに式 (6) の制約を課すと最小距離の下界は $2d_1 - 1$ となる [5]．表 1を見ると D_1 はこれらに対し十分大きくなっている．従って本稿で示したインタリーバの構成法および要素符号 C_1 の生成手法は有効であると考えられる．

5 むすび

本稿では H.Hagenauer らの提案したブロックターボ符号に対し，最小距離を大きくする要素符号の探索アルゴリズムおよびインタリーバの構成法を提案した．また結果的に得られるブロックターボ符号は従来より大きな最小距離を持つことを示した．

生成行列生成アルゴリズムでは，与えられた符号の生成行列に対するランダムな列置換を施して候補符号を生成した．しかし構成的に式 (14) を満足する符号を生成する手法を開発する必要がある．また，得られたブロックターボ符号 C_H^* に対する重み分布の導出なども今後の課題である．

謝辞

本研究の一部は文部科学省平成 12・13 年度科学研究費補助金 (課題番号 12875072) の助成による．

参考文献

[1] C.Berrou, A.Glavieux and P.Thitimajshima, "Near Shannon limit error-correcting Coding and Decoding: Turbo-codes(1), " in IEEE Int. Conf. Communications ICC'93, Vol.2/3, pp.1064-1071, May 1993.

[2] H.Hagenauer, E.Offer and L.Papke, "Iterative Decoding of Binary Block and Convolutional Codes, " IEEE Trans. Inform. Theory, Vol. IT-42, No.2, pp.429-445, March 1996.

[3] C.Heegard and S.B.Wicker, TURBO CODING, Kluwer Academic Publishers, 1999.

[4] R.M.Pyndiah, "Near-Optimum Decoding of Product Codes: Block Turbo Codes, " IEEE Trans. Commun. Vol.46, No.8, pp.1003-1010, August 1998.

[5] 小林 学, 松嶋 敏泰, 平澤 茂一, "ブロックターボ符号の生成行列と性能評価," 電子情報通信学会研究技術報告,IT2001-11, pp.1-6, July 2001.

The 24th Symposium on Information Theory
and Its Applications (SITA2001)
Kobe, Hyogo, Japan, Dec. 4 - 7, 2001

多端子情報理論に基づく分散協調問題について
Distributed Cooperative Problem based on Multi-Terminal Systems

吉田 隆弘* 松嶋 敏泰* 平澤 茂一*
Takahiro YOSHIDA Toshiyasu MATSUSHIMA Shigeichi HIRASAWA

Abstract— We consider the distributed cooperative systems for predictoin problem.In this paper, we apply multi-terminal systems to a distributed cooperative system and formulate a distributed cooperative problem based on Bayes decision theory. In the case that loss function is the logarithmic loss function, we shall show the Bayes rule minimizing Bayes risk and that Bayes risk of the Bayes rule is represented as mutual information.

Keywords— distributed cooperative problem, Bayes decision theory, source coding, side infomation

1 はじめに

コンピュータネットワーク上での分散協調問題とは，ネットワーク上で結合された複数の意思決定者 (エージェント) が個々に観測した情報と，エージェント間で情報のやりとりをすることによって得られた情報を利用して，各エージェントが全体のシステムとして良い出力が得られるように意思決定を行う問題である．このような分散強調の問題に対して，従来より様々なモデルと解法が研究されている [7][8]．本稿で扱う分散協調は，(1) 相関のある複数の情報を各エージェントがそれぞれ個別の情報として観測し，(2) エージェント間で情報交換を行い，(3) 得られた情報から各エージェントが個別に意思決定を行うものとする．このような分散協調の問題設定では，全ての情報を各エージェントが共有してしまえば，一人のエージェントによる基本的な決定問題と同等になってしまうので，エージェント間で交換される情報が重要な点であると考えられる．

一方，情報理論の分野において，複数の送・受信者の情報のやり取りを扱う多端子情報理論がある [1]．この分野における研究の一部では，情報源から発生する相関のある複数の情報を，複数の受信者が個別に受信した情報とは別に，補助的な情報や他の受信者から受け取った部分的な情報などを利用できるものとし，それらの情報を用いて復号を行うといった問題が扱われている [1]．また，多端子情報源符号化を情報源の確率分布パラメータを推定する問題や仮説検定に適用した研究もなされている [3][4]．パラメータ推定に関しては情報源から発生した情報をそのまま推定に用いるのではなく，通信路のレートの制約が満たされるように符号化した情報を用いて推定を行ったときの推定値の分散について研究が行われており，仮説検定に関しては同様に符号化した情報から検定を行ったときの誤り率について研究がなされている．

これらの問題設定における多端子モデルを，本稿で扱う分散協調問題と対応させてみると，各エージェントの決定は上述のモデルでの復号，パラメータ推定，および仮説検定を行うこととしてみることができ，各エージェントが観測する個別の情報とエージェント間の情報交換によって得られた情報は符号化された情報としてみることができる．

そこで本稿では，分散協調問題のモデルを多端子情報源符号化モデルとして捉え，二人のエージェント間で情報交換をして，意思決定するという分散協調問題をベイズ決定理論 [2] に基づいてモデル化し，ある損失に対して最適になるような各エージェントの決定を定式化する．本稿では，損失として対数損失を用いた場合のベイズ最適な決定と，その決定によるベイズリスクについての導出する．また，そのときに各エージェントがもう一方に与える情報の構成 (以下では符号化と呼ぶ) について考察を行う．

* 早稲田大学理工学部経営システム工学科，〒 169-8555 新宿区大久保 3-4-1, Dep. of Industrial and Management Systems Engineering, Waseda University, 3-4-1 Ohkubo, Shinjuku-ku, Tokyo 169-8555, E-mail:takahiro@matsu.mgmt.waseda.ac.jp

2 分散協調問題の定式化

2.1 モデル

本稿で扱う分散協調問題の基本モデルとして図 1 のような二人のエージェントの協調問題を考える．時点 t における情報源からの出力を確率変数 X_t, Y_t とし，それぞれ有限集合 $\mathcal{X}$, $\mathcal{Y}$ のなかに値をとる．さらに，$\mathcal{Y}$ の直積 $\mathcal{X}^n$, $\mathcal{Y}^n$ のなかに値をとる確率変数列を $X^t = X_1, X_2, \cdots, X_t$, $Y^t = Y_1, Y_2, \cdots, Y_t$ と表記する．本稿では確率変数 X_t, Y_t の同時確率分布が以下のようなパラメトリックな確率分布のクラス $\mathcal{P}$ に属していると仮定する．

$$\mathcal{P} = \left\{ p\left(x_t, y_t \mid x^{t-1}, y^{t-1}, \theta\right) \mid \theta \in \Theta \right\}. \qquad (1)$$

ここで，パラメータ $\theta \in \Theta$ は実数値ベクトルとし，確率分布のクラスは既知であるが，パラメータは未知であるとする．

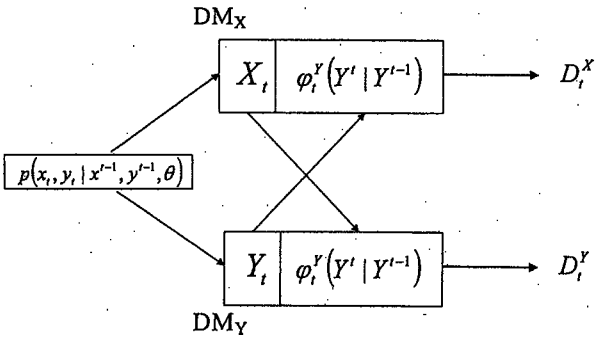

図 1 分散協調問題のモデル

各エージェント $\mathrm{DM}_X, \mathrm{DM}_Y$ は各時点において情報源からの出力を観測する．ただし，$\mathrm{DM}_X, \mathrm{DM}_Y$ は情報源からの出力の一方のみが観測できるものとする．すなわち，DM_X は X_t のみ，DM_Y は Y_t のみを各時点で観測でき，さらにエージェント DM_X, DM_Y は，もう一方のエージェントが観測した出力の部分的な情報 $\varphi_t^X\left(x_t \mid x^{t-1}\right)$, $\varphi_t^Y\left(y_t \mid y^{t-1}\right)$ をそれぞれ受け取ることができるものとする．ここで，

$$\varphi_t^X : \mathcal{X} \to \mathcal{U}_t \quad t = 1, 2, \cdots, n$$
$$\varphi_t^Y : \mathcal{Y} \to \mathcal{V}_t \quad t = 1, 2, \cdots, n$$

とした．また，$\mathcal{U}_t, \mathcal{V}_t$ は有限集合である．本稿では DM_X に対しては Y_t, DM_Y に対しては X_t を補助情報，上記の写像を符号化関数と呼ぶ．

各時点で交換する情報 $\varphi_t^X\left(x_t \mid x^{t-1}\right)$, $\varphi_t^Y\left(y_t \mid y^{t-1}\right)$ は，各エージェント DM_X,DM_Y がそれまでに観測した情報 x^{t-1}，y^{t-1} によって決定し，もう一方のエージェントに送るものとする．

3 決定理論に基づいたエージェントの推論

3.1 決定関数

前節で定義した分散協調モデルにおける各エージェントの推論を決定理論的にみると，各時点 t で各エージェントがそれまでに得られている情報対 $\left(x^{t-1},\varphi^Y\left(y^{t-1}\right)\right)$, $\left(y^{t-1},\varphi^X\left(x^{t-1}\right)\right)$ から，未知のある情報に対して逐次的に推測（決定）していく．本稿では，各エージェントに対する逐次決定関数を以下のように表記する．

$$D_t\left(x^t,y^{t-1},\varphi^X\left(x^{t-1}\right),\varphi^Y\left(y^{t-1}\right)\right) = \left\{D_t^X\left(x^t,\varphi^Y\left(y^{t-1}\right)\right),D_t^Y\left(y^t,\varphi^X\left(x^{t-1}\right)\right)\right\}. \quad (2)$$

ここで，

$$\varphi^X\left(x^t\right) = \left\{\varphi_1^X\left(x_1\right),\varphi_2^X\left(x_2 \mid x_1\right),\cdots,\varphi_t^X\left(x_t \mid x^{t-1}\right)\right\}, \quad (3)$$

$$\varphi^Y\left(y^t\right) = \left\{\varphi_1^Y\left(y_1\right),\varphi_2^Y\left(y_2 \mid y_1\right),\cdots,\varphi_t^Y\left(y_t \mid y^{t-1}\right)\right\} \quad (4)$$

とした．

この決定は $t-1$ 時点までの情報から t 時点の予測を行うことを毎期繰り返す逐次決定過程であるが，n 時点まで得られた情報 $\left(x^n,\varphi^Y\left(y^{n-1}\right)\right)$,$\left(y^n,\varphi^X\left(x^n\right)\right)$ から決定を行うことも考えられる．また，エージェントの決定に関しても，さまざまな決定関数を考えることができる．代表的な推論としては，分布のパラメータ θ の推定，次の時点において出現する x'，または y' の値の予測，x'，または y' の値ではなく，その分布関数の予測などがある．

3.2 損失関数と危険関数

損失は推論した結果と真の値との間のなんらかの距離を用いて評価されることが一般的である．推測の損失として良く用いられる距離としては，0-1 損失，二乗誤差損失，対数損失などがある [2]．ここで，式 (2) の逐次決定関数に対する損失関数を以下のように表記する．

$$L_t\left(D_t\left(x^t,y^{t-1},\varphi^X\left(x^{t-1}\right),\varphi^Y\left(y^{t-1}\right)\right),\theta\right) \quad (5)$$

また，時点 1 から時点 n までの累積損失関数を以下のように定義する．

$$L_n\left(D_{(n)}\left(x^n,y^n,\varphi^X\left(x^{n-1}\right),\varphi^Y\left(y^{n-1}\right)\right),\theta\right) = \sum_{t=1}^{n} L_t\left(D_t\left(x^t,y^{t-1},\varphi^X\left(x^{t-1}\right),\varphi^Y\left(y^{t-1}\right)\right),\theta\right). \quad (6)$$

ここで，時点 1 から時点 n までの決定の系列を

$$D_{(n)}\left(x^n,y^n,\varphi^X\left(x^{n-1}\right),\varphi^Y\left(y^{n-1}\right)\right) = \left\{D_{(n)}^X\left(x^n,\varphi^Y\left(y^{n-1}\right)\right),D_{(n)}^Y\left(y^n,\varphi^X\left(x^{n-1}\right)\right)\right\}, \quad (7)$$

$$D_{(n)}^X\left(x^n,\varphi^Y\left(y^{n-1}\right)\right) = \left\{D_1^X\left(x_1\right),\cdots,D_n^X\left(x^n,\varphi^Y\left(y^{n-1}\right)\right)\right\} \quad (8)$$

$$D_{(n)}^Y\left(y^n,\varphi^X\left(x^{n-1}\right)\right) = \left\{D_1^Y\left(y_1\right),\cdots,D_n^Y\left(y^n,\varphi^X\left(x^{n-1}\right)\right)\right\} \quad (9)$$

とした．

以上のように定義した損失関数に対する危険関数は以下のように定義される．

$$R_n\left(D_{(n)}\left(\varphi_{(n)}^Y,\varphi_{(n)}^X\right),\theta\right) = \sum_{x^n\in\mathcal{Y}^n}\sum_{y^n\in\mathcal{Y}^n} P\left(x^n,y^n \mid \theta\right) \times L_n\left(D_{(n)}\left(x^n,y^n,\varphi^X\left(x^{n-1}\right),\varphi^Y\left(y^{n-1}\right)\right),\theta\right). \quad (10)$$

ここで，$\varphi_{(n)}^X$,$\varphi_{(n)}^Y$ は，符号化関数の系列

$$\varphi_{(n)}^X = \left\{\varphi_1^X,\varphi_2^X,\cdots,\varphi_n^X\right\}, \quad (11)$$

$$\varphi_{(n)}^Y = \left\{\varphi_1^Y,\varphi_2^Y,\cdots,\varphi_n^Y\right\} \quad (12)$$

である．

3.3 ベイズ最適なエージェントの推論

決定理論において危険関数を最小化する代表的な基準として，ベイズ基準，maximin 基準，minimax 基準がある [2]．ここでは，前節で定義した危険関数をベイズ基準の下で最小化するエージェントの決定関数を求めることを考える．

危険関数に対するベイズリスクは以下のように定義される．

$$BR_n\left(D_{(n)}\left(\varphi_{(n)}^Y,\varphi_{(n)}^X\right)\right) = \int_\Theta R_n\left(D_{(n)}\left(\varphi_{(n)}^Y,\varphi_{(n)}^X\right),\theta\right)p\left(\theta\right)d\theta. \quad (13)$$

ここで，$p\left(\theta\right)$ はパラメータ θ の事前分布．また，符号化関数が与えられると，このベイズリスクを最小化する決定関数を求めることができ，このベイズ最適な決定関数はベイズ決定関数と呼ばれる．

4 予測を目的とした分散協調問題

4.1 ベイズ最適な予測

ここでは，各エージェントが，それまでに観測した情報を用いて，次の時点で観測される値の分布関数を予測するという決定について考える．すなわち，各時点 t で各エージェントがそれまでに得られている情報対 $\left(x^{t-1},\varphi^Y\left(y^{t-1}\right)\right)$, $\left(y^{t-1},\varphi^X\left(x^{t-1}\right)\right)$ から次の時点における X_t，Y_t の確率分布

$$p\left(x_t \mid x^{t-1},y^{t-1},\theta\right) = \sum_{y_{t+1}\in\mathcal{Y}} p\left(x_t,y_t \mid x^{t-1},y^{t-1},\theta\right), \quad (14)$$

$$p\left(y_t \mid x^{t-1},y^{t-1},\theta\right) = \sum_{x_s\in\mathcal{X}} p\left(x_t,y_t \mid x^{t-1},y^{t-1},\theta\right) \quad (15)$$

をそれぞれ推論する．したがって，各エージェントの決定関数は確率分布となる

また，この決定関数に対する損失関数，および累積損失関数を以下のような対数損失として定義する．

$$L_t\left(D_t\left(x^t,y^{t-1},\varphi^X\left(x^{t-1}\right),\varphi^Y\left(y^{t-1}\right)\right),\theta\right) = a\left\{\log\frac{p\left(x_t \mid x^{t-1},y^{t-1},\theta\right)}{D_t^X\left(x^t,\varphi^Y\left(y^{t-1}\right)\right)}\right\}$$

$$+b\left\{\log\frac{p\left(y_t\mid x^{t-1},y^{t-1},\theta\right)}{D_t^Y\left(y^t,\varphi^X\left(x^{t-1}\right)\right)}\right\},\quad(16)$$

$$\begin{aligned}
&L_n\left(D_{(n)}\left(x^n,y^n,\varphi^X\left(x^{n-1}\right),\varphi^Y\left(y^{n-1}\right)\right),\theta\right)\\
&\quad=a\sum_{t=1}^{n}\left\{\log\frac{p\left(x_t\mid x^{t-1},y^{t-1},\theta\right)}{D_t^X\left(x^t,\varphi^Y\left(y^{t-1}\right)\right)}\right\}\\
&\qquad+b\sum_{t=1}^{n}\left\{\log\frac{p\left(y_t\mid x^{t-1},y^{t-1},\theta\right)}{D_t^Y\left(y^t,\varphi^X\left(x^{t-1}\right)\right)}\right\}\\
&\quad=a\left\{\log\frac{p\left(x^n\mid y^{n-1},\theta\right)}{\prod_{t=1}^{n}D_t^X\left(x^t,\varphi^Y\left(y^{t-1}\right)\right)}\right\}\\
&\qquad+b\left\{\log\frac{p\left(y^n\mid x^{n-1},\theta\right)}{\prod_{t=1}^{n}D_t^Y\left(y^t,\varphi^X\left(x^{t-1}\right)\right)}\right\}.\quad(17)
\end{aligned}$$

ここで，a,b は非負の実数．

この損失関数に対するベイズリスクは，式 (10)，(13) より，

$$\begin{aligned}
&BR_n\left(D_{(n)}\left(\varphi_{(n)}^Y,\varphi_{(n)}^X\right)\right)\\
&\quad=aE_{p(X^n,Y^{n-1})}\left[\prod_{t=1}^{n}D_t^X\left(X^t,\varphi^Y\left(Y^{t-1}\right)\right)\right]\\
&\qquad+bE_{p(X^{n-1},Y^n)}\left[\prod_{t=1}^{n}D_t^Y\left(Y^t,\varphi^X\left(X^{t-1}\right)\right)\right]\\
&\qquad-aH\left(X^n\mid Y^{n-1},\Theta\right)-bH\left(Y^n\mid X^{n-1},\Theta\right)\quad(18)
\end{aligned}$$

となる．ここで，$E_{p(\cdot,\cdot)}$ は確率分布 $p(\cdot,\cdot)$ に関する期待値，$H(\cdot|\cdot)$ は条件付エントロピー [1] である．

以上のような予測において，DM_X に対する補助情報 y^{n-1} と DM_Y に対する補助情報 x^{n-1} がない場合，ユバーサル符号化 [1] の問題としてみることができ，このときの各エージェントは符号器の役割を果たす．また，式 (18) のベイズリスクは，補助情報付きのユニバーサル符号化でのベイズリスクと良く似た形となり [5]，このベイズリスクを最小にする決定関数も似た形で導出できる．

定理 4.1 任意の符号化関数 $\varphi_{(n)}^X,\varphi_{(n)}^Y$ に対し，式 (16) の損失関数に関する，時点 t でのベイズ決定関数は以下のように与えられる．

$$\begin{aligned}
&D_{t|op}\left(x^t,y^{t-1},\varphi^X\left(x^{t-1}\right),\varphi^Y\left(y^{t-1}\right)\right)\\
&\quad=\left\{D_{t|op}^X\left(x^t,\varphi^Y\left(y^{t-1}\right)\right),D_{t|op}^Y\left(y^t,\varphi^X\left(x^{t-1}\right)\right)\right\}\quad(19)
\end{aligned}$$

$$\begin{aligned}
&D_{t|op}^X\left(x^t,\varphi^Y\left(y^{t-1}\right)\right)\\
&\quad=\sum_{y^{t-1}}\int_\Theta p\left(x_t\mid x^{t-1},y^{t-1},\theta\right)\\
&\qquad\times p\left(\theta\mid x_{t-1},y^{t-1}\right)p\left(y^{t-1}\mid\varphi^Y\left(y^{t-1}\right)\right)d\theta,\quad(20)
\end{aligned}$$

$$\begin{aligned}
&D_{t|op}^Y\left(y^t,\varphi^X\left(x^{t-1}\right)\right)\\
&\quad=\sum_{x^{t-1}}\int_\Theta p\left(y_t\mid x^{t-1},y^{t-1},\theta\right)\\
&\qquad\times p\left(\theta\mid x_{t-1},y^{t-1}\right)p\left(x^{t-1}\mid\varphi^X\left(x^{t-1}\right)\right)d\theta.\quad(21)
\end{aligned}$$

□

4.2 ベイズ決定関数の性質

符号化関数の系列 $\varphi_{(n)}^X,\varphi_{(n)}^Y$ が与えられている場合，定理 4.1 のベイズ決定関数に関するベイズリスクは，

$$\begin{aligned}
&BR_n\left(D_{(n)}\left(\varphi_{(n)}^Y,\varphi_{(n)}^X\right)\right)\\
&\quad=a\sum_{x^n}\sum_{y^{n-1}}\int_\Theta p\left(x^n,y^{n-1},\theta\right)\\
&\qquad\times\log\frac{p\left(x^n\mid y^{n-1},\theta\right)}{p\left(x^n\mid\varphi^Y\left(y^{n-1}\right)\right)}d\theta\\
&\qquad+b\sum_{x^{n-1}}\sum_{y^n}\int_\Theta p\left(x^n,y^{n-1},\theta\right)\\
&\qquad\times\log\frac{p\left(y^n\mid x^{n-1},\theta\right)}{p\left(y^n\mid\varphi^X\left(x^{n-1}\right)\right)}d\theta\quad(22)
\end{aligned}$$

と書ける．ここで，式 (22) の右辺第一項がエージェント DM_X の決定に対するベイズリスクとなり，これを $BR_n^X(\varphi_{(n)}^Y)$ とおくと，

$$\begin{aligned}
&BR_n^X\left(\varphi_{(n)}^Y\right)\\
&\quad=a\sum_{x^n}\sum_{y^{n-1}}\int_\Theta p\left(x^n,y^{n-1},\theta\right)\\
&\qquad\times\log\frac{p\left(x^n\mid y^{n-1},\theta\right)p\left(x^n\mid y^{n-1}\right)p\left(x^n\right)}{p\left(x^n\mid y^{n-1}\right)p\left(x^n\right)p\left(x^n\mid\varphi^Y\left(y^{n-1}\right)\right)}d\theta\\
&\quad=aI\left(X^n;\Theta\mid Y^{n-1}\right)+aI\left(X^n;Y^{n-1}\right)\\
&\qquad-aI\left(X^n;\varphi^Y\left(Y^{n-1}\right)\right)\quad(23)
\end{aligned}$$

となる．また，式 (22) の右辺第二項はエージェント DM_Y の決定に対するベイズリスクとなり，これを $BR_n^Y(\varphi_{(n)}^X)$ とおき，上式と同様の展開をすると，

$$\begin{aligned}
&BR_n^Y\left(\varphi_{(n)}^X\right)=bI\left(Y^n;\Theta\mid X^{n-1}\right)+bI\left(Y^n;X^{n-1}\right)\\
&\qquad-bI\left(Y^n;\varphi^X\left(X^{n-1}\right)\right)\quad(24)
\end{aligned}$$

となり，本稿で定義している分散協調問題において符号化関数が与えられた場合，各エージェントがベイズ最適な決定をしたときのベイズリスクを 6 つの相互情報量の和として表現できることが示せた．

ここで，エージェント DM_X の決定に対するベイズリスク $BR_n^X(\varphi_{(n)}^Y)$ について考える．右辺第一項の値は条件付相互情報量となっており，この量は Y^{n-1} の情報全てを DM_X が観測した場合でのベイズ決定関数によるベイズリスクとみなすことができる．したがって，エージェント DM_X が Y^{n-1} の情報全てを知っていたとしても残ってしまうベイズリスクが第一項の条件付相互情報量となり，DM_X が Y^{n-1} の情報を完全に受け取ることができなかったために生じるベイズリスクが残りの第二項，第三項によって表現されている．第三項については，$\varphi^Y\left(Y^{n-1}\right)$ の中に Y^{n-1} の情報が多く含まれているほど大きい値となり，第二項に近い値をとるのでベイズリスクが減少する．よって

$$\begin{aligned}
&aI\left(X^n;\Theta\mid Y^{n-1}\right)\le BR_n^X\\
&\qquad\le a\left(I\left(X^n;\Theta\mid Y^{n-1}\right)+I\left(X^n;Y^{n-1}\right)\right)\quad(25)
\end{aligned}$$

が常に成立している．この右辺第二項の相互情報量は，DM_X が観測する X^n と補助情報である Y^{n-1} との相関の高さを示す量である．また，DM_X が Y^{n-1} の情報を全く受け取ることができない場合というのは，Y^{n-1} に含まれている X^n の情報を推論に用いることができないためベイズリスクが増加し，さらに，Y^{n-1} と X^n の相関が高いほど多くの情報が得られていないことになるので，Y^{n-1} の情報を全く受け取ることができないことによる損失がより大きくなりベイズリスクの増加に大きく影響することになる．これが上式の右辺第二項の意味である．式 (24) 右辺の 3 つの相互情報量についても $BR_n^X(\varphi_{(n)}^Y)$ と同様の意味をもち，

$$bI\left(Y^n;\Theta \mid X^{n-1}\right) \le BR_n^Y \le b\left(I\left(Y^n;\Theta \mid X^{n-1}\right)+I\left(Y^n;X^{n-1}\right)\right) \quad (26)$$

が成立している．

4.3 符号化関数の構成

前節では，符号化関数が与えられているもとでのベイズリスクについて述べた．本節では，定理 4.1 のベイズ決定関数によるベイズリスクを最小にする符号化関数について考える．ここでは，符号化関数に対する制約を以下で与える．

$$\log|\cup_{t=1}^n \mathcal{U}_t| \le nR^X, \quad \log|\cup_{t=1}^n \mathcal{V}_t| \le nR^Y \quad (27)$$

とする．また，符号化関数 $\varphi_{(n)}^X$,$\varphi_{(n)}^Y$ を，試験通信路

$$Q_{n-1}^X\left(u_{n-1} \mid x^{n-1}\right), Q_{n-1}^Y\left(v_{n-1} \mid y^{n-1}\right) \quad (28)$$

として表す．ここで，

$$u_{n-1} \in \mathcal{U}^{(n-1)} = \cup_{t=1}^{n-1}\mathcal{U}_t, \quad (29)$$

$$v_{n-1} \in \mathcal{V}^{(n-1)} = \cup_{t=1}^{n-1}\mathcal{V}_t. \quad (30)$$

式 (16) より，各エージェントの決定に関する損失は線形となっているので，$BR^X(\varphi_{(n)}^Y)$,$BR^Y(\varphi_{(n)}^X)$ のそれぞれについて最小になる符号化関数を選択すれば，全体のベイズリスクも最小化されるので，式 (23) についてのみ考えることとする．式 (23) は以下のように書き換えられる．

$$\begin{aligned} BR_n^X\left(Q_{n-1}^Y\right) &= a\sum_{x^n}\sum_{y^{n-1}}\sum_{v_{n-1}}\int_\Theta p\left(x^n,y^{n-1},v_{n-1},\theta\right) \\ &\quad \times \log\frac{p\left(x^n \mid y^{n-1},\theta\right)p\left(x^n \mid y^{n-1}\right)}{p\left(x^n \mid y^{n-1}\right)p\left(x^n \mid v_{n-1}\right)}d\theta \\ &= aI\left(X^n;\Theta \mid Y^{n-1}\right)+aI\left(X^n;Y^{n-1}\right) \\ &\quad -aI\left(X^n;V^{n-1}\right), \end{aligned} \quad (31)$$

$$\begin{aligned} &p\left(x^n,y^{n-1},v_{n-1},\theta\right) \\ &\quad = p\left(x^n \mid y^{n-1},v_{n-1},\theta\right)p\left(y^{n-1} \mid \theta\right) \\ &\qquad \times Q_{n-1}^Y\left(v_{n-1} \mid y^{n-1}\right)p\left(\theta\right) \end{aligned} \quad (32)$$

試験通信路に依存した量は第三項の相互情報量のみなので，式 (27) の制約のもと，この量をを最大化することで $BR_n^X\left(Q_{n-1}^Y\right)$ が最小になる．したがって，試験通信路 $Q_{n-1}^Y\left(v_{n-1} \mid y^{n-1}\right)$ が，

$$I\left(Y^{n-1};V^{n-1}\right) \le R^Y \quad (33)$$

を満たしている同時確率分布 $p(x^n,y^{n-1},v_{n-1})$ のクラスから，$I(X^n;V^{n-1})$ を最大にする条件付確率分布 $p(v_{n-1}$ $x^n)$ を持つような同時確率分布を探すことで，式 (27) の制約下で最適な符号化関数が構成できる．$BR_n^Y\left(Q_{n-1}^X\right)$ についても，同様のことがいえる．この試験通信路は，通信路容量，あるいは Rate-distortion 関数の既存の算法 [1] を応用することで構成できると考えられる．

5 まとめ

本稿では，多端子情報理論とベイズ決定理論に基づいて分散協調問題をモデル化し，各エージェントの推論を定式化した．また，対数損失による損失関数に対して，ベイズ最適な決定関数を導出し，そのときのベイズリスクが相互情報量と条件付相互情報量で表現できることを示した．さらに，符号化関数の構成について考察し，相互情報量を最大化することで定められることを示したが，具体的な符号化関数の構成について，損失関数の一般化などが今後の課題として考えられる．

6 謝辞

本研究を行うにあたり，数多くの御助言，御支援を賜りました，早稲田大学松嶋研究室　野村亮氏並びに松嶋研究室内各氏に心より感謝申し上げます．なお，本研究の一部は文部省科学研究費基盤 (C)(No.12650400)，早稲田大学特定課題研究助成費 (2001A-570) の援助による．

参考文献

[1] T.M.Cover and J.A.Thomas, "Elements of information theory," John Wiley & Sons, New York, 1991.

[2] J.O.Berger, "statistical decision theory and Bayesian analysis," Springer-Verlag, New York, 1985.

[3] T.H.Han and S.Amari, "Parameter estimation with multiterminal data compression," *IEEE Trans. Inf. Theory*, vol.41, no.6, pp.1802-1833, Nov. 1995.

[4] S.Amari and T.H.Han, "Statistical inference under multiterminal rate restrictions: A differential geometrical approach," *IEEE Trans. Inf. Theory*, vol.35, no.2, pp.217-227, Mar. 1989.

[5] 松嶋敏泰, "帰納・演繹推論と予測 -決定理論による学習モデル-," 1998 年情報論的学習理論ワークショップ予稿集, pp.1-8, 1998.

[6] T.Matsushima, H.Inazumi and S.Hirasawa,"A Class of Distotionless Codes Designed by Bayes Decision Theory," *IEEE Trans. Inf. Theory*, vol.37, no.5, pp.1288-1293, Sept. 1991.

[7] K.Yamanishi, "Distributed Cooperative Bayesian Learning Straregies," Information and Computation 150, pp.22-56, 1999.

[8] 横尾真, E.H.Durfee, 石田亨, 桑原和宏, "分散制約充足による分散協調問題解決の定式化とその解法," 電子情報通信学会論文誌 D-1, vol.J75-D-1, no.8, pp.704-713, 1992.

The 25th Symposium on Information Theory
and Its Applications (SITA2002)
Ikaho, Gunma, Japan, Dec. 10–13, 2002

拡張された階層モデルにおける予測アルゴリズムについて
Prediction Algorithm for Extended Hierarchical Models

須子 統太* 野村 亮* 松嶋 敏泰*
Tota Suko Ryo Nomura Toshiyasu Matsushima

Abstract— In the field of sources coding,the Bayes coding algorithm which codes by calculating a predictive distribution is devised. Moreover, in the field of signal processing, the algorithm which estimate the base of a signal by applying Bayes coding algorithm is devised. In this paper, we extend the model class by regarding previous works systematically as a prediction problem from data. Moreover, We show the Bayes optimal prediction algorithm in the extended model class.

Keywords— Bayes decision theory, hierarchical model, prediction problem

1 はじめに

与えられたデータから次に発生するデータを予測する問題は，従来から様々な分野において研究がなされている．情報源の分布のクラスのみを仮定し，そのパラメータに関しては未知な場合を扱うユニバーサル符号化において，符号化確率に予測分布を用いて符号化を行う研究が行われている [1][3]．また信号処理の分野では，雑音を含んだ信号からもとの信号を推定する為に予測分布を用いる研究が行われている [2]．これらは共に予測分布を計算する予測問題と捉えることができる．

複数のモデルを考慮した予測問題において，ベイズ最適な予測を行う為には定義されたモデルクラスに含まれる全てのモデルにおける予測分布の重み付け和を計算する必要がある．その為，計算量がモデルの数のオーダで増えることになる．[1] では，FSMX 情報源に対して効率的に予測分布を計算するベイズ符号アルゴリズムを提案している．また，[2] ではウェーブレットパケット木と呼ばれるモデルクラスが FSMX 情報源と類似の構造を持つことに注目し，ベイズ符号アルゴリズムを改良することにより効率的な予測分布計算アルゴリズムを提案している．

本研究では複数のモデルを考慮した予測問題に対して，従来扱われていたモデルクラスを拡張したモデルクラスを定義する．また，拡張したモデルクラスに対して効率的な予測分布計算アルゴリズムを提案する．

2 問題設定

2.1 予測問題

入力と出力の n 個の組 $z^n = \{x^n, y^n\}$ が得られたとする．但し，$x^n = x_1, x_2, ..., x_n$, $y^n = y_1, y_2, ..., y_n$, $x \in X$, $y \in Y$ とする．† このとき，$n+1$ 番目の入力 x_{n+1} が与えられたもとで x_{n+1} に対応する出力 y_{n+1} を予測することを考える．

また，x,y は $P(x|\nu)$,$P(y|x,\theta_m,m)$ にそれぞれ従うとする．但し,$\nu \in N$ を x の分布のパラメータ，$m \in M$ をモデル,$\theta_m \in \Theta_m$ をモデル m のもとでの y の分布のパラメータとする．

2.2 ベイズ最適な予測

予測に対する損失を $Loss$ で表す．損失は以下で示されるような様々な損失が考えられる．$\hat{y}$ を y の予測値，$\hat{P}(y|x)$ を $P(y|x)$ の予測値とする．

- 二乗誤差損失

$$Loss_1 = (y_{n+1} - \hat{y}_{n+1})^2. \tag{1}$$

- 0-1 損失

$$Loss_2 = \begin{cases} 0 & y_{n+1} = \hat{y}_{n+1} \\ 1 & y_{n+1} \neq \hat{y}_{n+1} \end{cases}. \tag{2}$$

- 対数損失

$$Loss_3 = \log P(y_{n+1}|x_{n+1}) - \log \hat{P}(y_{n+1}|x_{n+1}). \tag{3}$$

また，データによる損失の期待値である危険関数を以下の式で定義する．

$$Risk = \sum_{Y^n} \sum_{X^n} Loss \times P(y^n|x^n,\theta_m,m)P(x^n|\nu). \tag{4}$$

更に，事前分布で期待値をとったベイズ危険関数を以下の式で定義する．

$$BR = \sum_M \int_N \int_{\Theta_m} Risk \times P(m)P(\theta_m|m)P(\nu)d\theta_m d\nu. \tag{5}$$

ベイズ最適な予測は (5) 式を最小にする予測である．このとき，ベイズ最適な予測は各損失関数によって以下で求められる．

- $Loss_1$ に対する最適な予測

$$\hat{y}^*_{n+1} = \sum_Y y_{n+1} \sum_M \int_{\Theta_m} P(y_{n+1}|x_{n+1},\theta_m,m) P(\theta_m|z^n)P(m|z^n)d\theta_m. \tag{6}$$

- $Loss_2$ に対する最適な予測

$$\hat{y}^*_{n+1} = \arg\max_{y_{n+1}} \sum_M \int_{\Theta_m} P(y_{n+1}|x_{n+1},\theta_m,m) P(\theta_m|z^n)P(m|z^n)d\theta_m. \tag{7}$$

- $Loss_3$ に対する最適な予測

$$\hat{P}^*(y_{n+1}|x_{n+1},z^n) = \sum_M \int_{\Theta_m} P(y_{n+1}|x_{n+1},\theta_m,m) P(\theta_m|z^n)P(m|z^n)d\theta_m. \tag{8}$$

* 〒 169-8555 東京都新宿区大久保 3-4-1 早稲田大学理工学部経営システム工学科. Dept. of Industrial & Management Systems Engineering, School of Science and Engineering, Ookubo 3-4-1, Shinjyukuku, Tokyo, 169-8555 Japan. E-mail: suko@matsu.mgmt.waseda.ac.jp

† 本研究では特に断らないかぎり x および y を離散値として扱うが，連続値としても同様の議論が可能である．

つまり (8) 式で表される予測分布を計算すれば，最適な予測が可能になる．パラメータの事前分布に自然共役な事前分布を仮定すると，パラメータ空間における積分計算にかかる計算量はデータ数 n の線形オーダとすることができる．他方，モデルに対する期待値を求める計算に必要な計算量は $O(|M|)$ となり，モデル数が多い場合この部分の計算量が全体の計算量の主要項となる．本研究では，このモデルに対する期待値計算の効率化を目的とする．

3 効率的アルゴリズム

3.1 モデルクラスの定義

x の空間 X を分割することを考える．この時 X の既約空間を $c_a \in C$ とする．但し $a = 1, \cdots, |C|$，$|C|$ は有限である．また，X から C への写像を f とする．例えば，$x \in \{0,1\}^2$ の時 $C = \{c_1 = 00, c_2 = 10, c_3 = 01, c_4 = 11\}$ となり，x が離散値の場合 c_a は x そのもので表現される．また x が連続値の場合は，例えば $x \in [0,1)$ の時 $C = \{c_a | a = 1,2,3,4, c_a = f(x), 0.25(1-a) < x \leq 0.25a\}$ のように，実問題ごとに予測者によって定められた f によって c_a は表現される．

次に，c_a の集合によって表現される状態を $s \in S$ とする．s は C の部分集合になっている．また，

$$\sigma_a = \begin{cases} 0 & c_a \notin s \\ 1 & c_a \in s \end{cases}. \tag{9}$$

とする．これを用いて s には $s_{\sigma_1\sigma_2\cdots\sigma_{|C|}}$ とインデクスをつける．以降 $\sigma_1\sigma_2\cdots\sigma_{|C|}$ を二進数ととらえ，アルファベット小文字一文字で表す．

この時，モデルクラス M を以下の式を満たす m の集合として定義する．

$$m = \{s_v, s_w, \cdots | \forall k,l \in \{0,1\}^{|C|}, k \neq l, s_k \neq \phi, s_k \cap s_l = \phi, \cup s_k = C\}. \tag{10}$$

$\theta_s \in \Theta_s$ を状態 s のもとでの y の分布のパラメータとし，$S(m)$ は m に含まれる s の集合とする時，モデル m におけるパラメータを，$\theta_m = \{\theta_s | s \in S(m)\}$，と θ_s の集合として表すとする．また，$s(m,x)$ を $S(m)$ の中で $f(x) = c$ を含む s とした時，y の分布を以下で定義する．

$$P(y|x,\theta_m,m) = P(y|x,\theta_{s(m,x)},s(m,x)). \tag{11}$$

つまり，モデル m のもとでの y の分布は，入力 x によって一意に定まる状態 $s(x,m)$ のパラメータ $\theta_{s(x,m)}$ にのみ依存するものとする．

また，s の事前分布を以下で定義する．

$$P(s) = \sum_{\{m|s\in S(m)\}} P(m). \tag{12}$$

更に，S の中で，$f(x) = c$ を含む s の集合を，

$$S(x) = \{s_{t_1}, s_{t_2}, \cdots s_{t_h}\}, \tag{13}$$

とする．但し $t_1 < t_2 < \cdots < t_h$, である．このとき，次式が満たされる．

$$\sum_{i=1}^{h} P(s_{t_i}) = 1, \tag{14}$$

尚，複数のモデルに同一の s が含まれてくるが，s のパラメータ θ_s の事前分布 $P(\theta_s)$ は同一の s であれば全て等しいと仮定する．

3.2 効率的アルゴリズム

$P(s_{t_i}|z^n)$ を次式のように表現する．

$$P(s_{t_i}|z^n) = q(s_{t_i}|z^n) \prod_{j>i} (1 - q(s_{t_j}|z^n)). \tag{15}$$

【効率的アルゴリズム】

step-1. x_{n+1} を受け取る．

step-2. $f(x_{n+1}) = c$ から $S(x_{n+1})$ を求める．

step-3. $S(x_{n+1})$ に含まれる全ての s について以下の再帰計算を行う．

$$q(y_{n+1}|x_{n+1},z^n,s_{t_i}) = \begin{cases} P^S(y_{n+1}|x_{n+1},z^n,s_{t_i}) & i = 1 \\ (*) & \text{otherwise} \end{cases}. \tag{16}$$

$$(*) = q(s_{t_i}|z^n)P^S(y_{n+1}|x_{n+1},z^n,s_{t_i}) + (1 - q(s_{t_i}|z^n))q(y_{n+1}|x_{n+1},z^n,s_{t_{i-1}}). \tag{17}$$

但し，

$$P^S(y_{n+1}|x_{n+1},z^n,s_{t_i}) = \int_{\Theta_s} P(y_{n+1}|x_{n+1},\theta_{s_{t_i}},s_{t_i}) P(\theta_{s_{t_i}}|z^n) d\theta_{s_{t_i}}. \tag{18}$$

step-4. $q(y_{n+1}|x_{n+1},z^n,s_{t_h})$ を予測分布として予測を行う．

step-5. $S(x_{n+1})$ に含まれる全ての s について，$q(s_{t_i}|z^n)$ を以下の式で更新する．

$$q(s_{t_i}|z^{n+1}) = \frac{q(s_{t_i}|z^n)P^S(y_{n+1}|x_{n+1},z^n,s_{t_i})}{(**)}. \tag{19}$$

$$(**) = q(s_{t_i}|z^n)P^S(y_{n+1}|x_{n+1},z^n,s_{t_i}) + (1 - q(s_{t_i}|z^n))q(y_{n+1}|x_{n+1},z^n,s_{t_{i-1}}). \tag{20}$$

step-6. step-1 へ戻る．

上記のアルゴリズムを用いた場合，モデルに対する期待値を計算するのに必要な計算量は $O(|S(x_{n+1})|)$ となる．また，上記アルゴリズムによる予測はベイズ最適な予測になっている．(付録参照)

4 考察

4.1 従来研究との関係

4.1.1 ユニバーサル無歪データ圧縮

[1] では，長さ r の系列のもとで $r+1$ 番目のシンボルの予測分布を符号化確率として符号化を行っている．深さ D の FSMX 情報源を仮定した場合，0 次から D 次までのマルコフモデルを含むモデルクラスを仮定していることになる．つまり，入力 x を $r-D$ 番目から r 番目までの系列，出力 y を $r+1$ 番目のシンボルと考えると，2.1 で示した問題設定と等価な問題であると言える．また，損失は (3) 式の $Loss_3$ を用いている．

モデルクラスである FSMX 情報源は情報源アルファベット数を K とした時，s をノードとした K 進木で表現することができる．根ノードは $s = C$ となる s，葉ノードは $s = c_a$ となる s を用いていて，親ノードは子ノード

の和集合で表現される．各モデルはノードを表している s の中から，(10) 式を満たしている s の集合で表現されている．M が (10) 式を満たす全ての m を含む（以降，フルモデルクラスと呼ぶ）場合，C の任意の分割の組み合わせ数だけのモデルを含んでいるのに対し，FSMX 情報源ではフルモデルクラスのうち上記のような木表現が可能なモデルのみを用いていると言える．

例えば，$D=2$ で情報源アルファベットが $\{0,1\}$ の場合，$C=\{c_1=00, c_2=10, c_3=01, c_4=11\}$ の 4 状態を考える．このとき，

$$
\begin{aligned}
m_1 &= \{s_{1111}\},\\
m_2 &= \{s_{1100}, s_{0011}\},\\
m_3 &= \{s_{1000}, s_{0100}, s_{0011}\},\\
m_4 &= \{s_{1100}, s_{0010}, s_{0001}\},\\
m_5 &= \{s_{1000}, s_{0100}, s_{0010}, s_{0001}\},
\end{aligned}
$$

の 5 種類のモデルのみを含む M が FSMX 情報源となる．つまり，r 番目のシンボルは $r-1$ 番目のシンボルに，$r-1$ 番目のシンボルは $r-2$ 番目のシンボルに依存するという時系列的な制約があるモデルクラスであると言える．

また，FSMX 情報源では $P(y|x,\theta_m,m)$ の分布に多項分布を仮定している．

4.1.2 信号推定

[2] では，ウェーブレットパケット基底を用いて，雑音が混入した離散の観測信号からの未知信号推定を行っている．これは，基底と未知信号の変換係数を未知パラメータとした推定問題である．しかしながら，x を観測信号，y を未知パラメータと考えたとき，z^0 のもとで x_1 から y_1 を予測する予測問題ととらえることで，2.1 で示した問題設定と等価であると言える．また，損失としては (1) 式の $Loss_1$ を用いている．

モデルクラスであるウェーブレットパケット木は，二進木で表現され各ノードは正規直交基底を表している．また，各ノードは子ノードの直和になっている．特に，根ノードは離散信号の空間 X に対応しており，葉ノードは既約空間 c_a に対応している．更に，各モデルは X を直和分割した正規直交基底の集合によって表現されている．つまり，本研究の表記を用いると前述の FSMX 情報源と全く等価なモデルクラスであると言える．但し，FSMX 情報源では $P(y|x,\theta_m,m)$ の分布に多項分布を仮定していたのに対し，ウェーブレットパケット木では正規分布を仮定しているという違いがある．

4.2 計算量

前述したように，効率的アルゴリズムを用いることで，モデルに対する期待値の計算に必要な計算量を $O(|M|)$ から $O(|S(x_{n+1}|)$ とすることができる．3.1 で定義したモデルクラスにおいて，$|M|$ は $|C|$ に依存してくる．フルモデルクラスの場合，

$$
|M| = \sum_{v=1}^{|C|}\sum_{w=0}^{v-1}\binom{v}{w}\frac{(-1)^w(v-w)^{|C|}}{u!} \tag{21}
$$

同様に，$|S(x_{n+1})|$ も $|C|$ に依存してくる．フルモデルクラスの場合，

$$
|S(x_{n+1})| = 2^{|C|-1}, \tag{22}
$$

となり，効率的に計算されていることが分かる．

また，[1]，[2] で扱われているモデルクラスの場合，$|C|=2^D$ となり，$|M|$ は D に依存してくる．ここで，M_D を直前の D 個のシンボルに依存する場合のモデルクラスする．このとき，

$$
\begin{aligned}
|M_D| &= 1+|M_{D-1}|^2\\
&> 2^{2^D} = 2^{|C|}, \qquad (23)\\
|S(x_{n+1})| &= \log_2|C|+1, \qquad (24)
\end{aligned}
$$

となっている．

4.3 その他の応用分野

従来研究では木表現が可能なモデルクラスを仮定していたのに対し，3.1 で定義したモデルクラスでは木表現ができない場合のモデルクラスも含まれてくる．そのため従来扱われていなかった，より様々な問題設定に適応可能であると考えられる．応用可能なその他の問題としては，テキスト分類，パターン認識などが考えられる．

5 まとめ

本研究では複数のモデルを考慮した予測問題に対して，従来扱われていたモデルクラスを拡張したモデルクラスを定義した．また，拡張したモデルクラスのもとで効率的に予測分布を計算するアルゴリズムを提案した．それにより，従来扱われていなかった他の応用分野において，効率的に予測分布を計算可能となった．

今回はモデルの事前分布の設定方法について特に言及しなかった．事前分布の設定方法によっては，予測性能をあまり劣化させずメモリ量を削減する近似ベイズ最適アルゴリズムが構成可能であると考えられる．これについては今後の課題としたい．

謝辞

本研究を行うにあたり，数多くの御助言，御支援を賜りました，浮田善文氏，吉田隆弘氏，並びに松嶋研究室の各氏に感謝致します．なお，本研究の一部は文部省科学研究費基盤 (C)(No.12650400)，早稲田大学特定課題研究助成費 (2001A-570) の援助による．

文献

[1] T. Matsushima, and S. Hirasawa, "A bayes coding using context tree." In Proc. Int. Symp. on. Inf. Theory, page 386, 1994.

[2] 北原正樹, 野村亮, 松嶋敏泰, "ウェーブレットパケット基底を用いた信号推定におけるベイズ決定理論の適用に関する一考察." 信学論 (A), vol.J85-A, No.5, page 584, 2002.

[3] T. Matsushima, H. Inazumi and S. Hirasawa, "A Class of Distortionless Codes Designed by Bayes Decision Theory" IEEE Trans. Inf. Theory, vol.37, No.5, page 1288, 1991.

[4] 松嶋敏泰, "帰納・演繹推論と予測-決定理論による学習モデル-." 1998 年情報論的学習理論ワークショップ,page 1, 1998.

[5] Frans M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, "Context Weighting for General Finite-Context Sources." IEEE Trans. Inf. Theory, vol.42, No.5, page 1514, 1996.

付録

効率的アルゴリズムによる予測がベイズ最適になっていることを以下で証明する.

証明

アルゴリズムによって計算された予測分布が (8) 式と等しければ, ベイズ最適性を証明できる.

まず, (16) 式が (8) 式を計算していることを示す.

(16) 式を展開すると,

$$\begin{aligned}
&q(y_{n+1}|x_{n+1},z^n,s_{t_h})\\
&\quad=P^S(y_{n+1}|x_{n+1},z^n,s_{t_h})q(s_{t_h}|z^n)+\cdots\\
&\quad\quad+P^S(y_{n+1}|x_{n+1},z^n,s_{t_i})q(s_{t_i}|z^n)\\
&\quad\quad\times\textstyle\prod_{j>i}(1-q(s_{t_j}|z^n))+\cdots\\
&\quad\quad+P^S(y_{n+1}|x_{n+1},z^n,s_{t_1})\textstyle\prod_k(1-q(s_{t_k}|z^n)).
\end{aligned}\tag{25}$$

このとき, $q(s_{t_1}|z^n)=1$, および (15) 式より,

$$\begin{aligned}
&q(y_{n+1}|x_{n+1},z^n,s_{t_h})\\
&\quad=\textstyle\sum_i P^S(y_{n+1}|x_{n+1},z^n,s_{t_i})q(s_{t_i}|z^n)\\
&\quad\quad\times\textstyle\prod_{j>i}(1-q(s_{t_j}|z^n))\\
&\quad=\textstyle\sum_i P(s_{t_i}|z^n)P^S(y_{n+1}|x_{n+1},z^n,s_{t_i})\\
&\quad=\textstyle\sum_{s\in S(x_{n+1})}\int_{\Theta_s}P(y_{n+1}|x_{n+1},\theta_s,s)P(\theta_s|z^n)d\theta_s\\
&\quad\quad\times P(s|z^n).
\end{aligned}\tag{26}$$

また, (8) 式は (10),(11),(12) 式より,

$$\begin{aligned}
&\sum_M\int_{\Theta_m}P(y_{n+1}|x_{n+1},\theta_m,m)P(\theta_m|z^n)P(m|z^n)d\theta_m\\
&\quad=\textstyle\sum_{s\in S(x_{n+1})}\int_{\Theta_s}P(y_{n+1}|x_{n+1},\theta_s,s)P(\theta_s|z^n)d\theta_s\\
&\quad\quad\times\textstyle\sum_{m|s\in S_m}P(m|z^n)\\
&\quad=\textstyle\sum_{s\in S(x_{n+1})}\int_{\Theta_s}P(y_{n+1}|x_{n+1},\theta_s,s)P(\theta_s|z^n)d\theta_s\\
&\quad\quad\times P(s|z^n).
\end{aligned}\tag{27}$$

よって, (26), (27) 式より,

$$\begin{aligned}
&\sum_M\int_{\Theta_m}P(y_{n+1}|x_{n+1},\theta_m,m)P(\theta_m|z^n)P(m|z^n)d\theta_m\\
&\quad=q(y_{n+1}|x_{n+1},z^n,s_{t_h}).
\end{aligned}\tag{28}$$

以上より, (16) 式が (8) 式を計算していることが示された.

また, アルゴリズムでは逐次的に予測を行なため, $n+1$ 時点での予測には n 時点までのモデルの事後分布を事前分布として用いる. そのため, (19) 式による $q(s_{t_i}|z^{n+1})$ の計算が, 正しく $P(m|z^{n+1})$ を計算している必要がある.

このとき, $P(m)=P(m|z^0)$,$P(s)=P(s|z^0)$ であることから, ベイズの定理より,

$$\begin{aligned}
P(s|z^1)&=\frac{P(s|z^0)P^S(y_1|x_1,z^0,s)}{\sum_{s'\in S}P(s'|z^0)P^S(y_1|x_1,z^0,s')}\\
&=P(s|z^0)P^S(y_1|x_1,z^0,s)\\
&\quad/\{\textstyle\sum_{s'\in S}P(s'|z^0)P^S(y_1|x_1,z^0,s')\\
&\quad+\textstyle\sum_{s''\notin S}P(s''|z^0)P^S(y_1|x_1,z^0,s'')\}.
\end{aligned}\tag{29}$$

また, (11), (12) 式より,

$$\begin{aligned}
&P(s|z^1)\\
&\quad=\frac{\sum_{\{m|s\in S(m)\}}P(m|z^0)P^S(y_1|x_1,z^0,s)}{\sum_{s'\in S}\sum_{\{m|s'\in S(m)\}}P(m|z^0)P^S(y_1|x_1,z^0,s')}\\
&\quad=\frac{\sum_{\{m|s\in S(m)\}}P(m|z^0)P^m(y_1|x_1,z^0,m)}{\sum_{m'\in M}P(m|z^0)P^m(y_1|x_1,z^0,m')}.
\end{aligned}\tag{30}$$

但し,

$$\begin{aligned}
&P^m(y_{n+1}|x_{n+1},z^n,m)\\
&\quad=\textstyle\int_{\Theta_m}P(y_{n+1}|x_{n+1},\theta_m,m)P(\theta_m|m,z^n)d\theta_m.
\end{aligned}\tag{31}$$

とする. よってベイズの定理より,

$$P(s|z^1)=\sum_{\{m|s\in S\}}P(m|z^1).\tag{32}$$

が成り立つ. $P(s|z^k)=\sum_{\{m|s\in S\}}P(m|z^k)$ が成立すると仮定すると, 同様に, $P(s|z^{k+1})=\sum_{\{m|s\in S\}}P(m|z^{k+1})$ が成立する. よって,

$$P(s|z^{n+1})=\sum_{\{m|s\in S\}}P(m|z^{n+1}).\tag{33}$$

つまり (33) 式より, $q(s_{t_i}|z^{n+1})$ の計算が $P(s_{t_i}|z^{n+1})$ を計算していることが示されれば, $P(m|z^{n+1})$ を計算していることを示せる. すなわち (15) 式より,

$$P(s_{t_i}|z^{n+1})=q(s_{t_i}|z^{n+1})\prod_{j>i}(1-q(s_{t_j}|z^{n+1})),\tag{34}$$

が示されれば良い.

まず, ベイズの定理より (34) 式左辺は,

$$P(s_{t_i}|z^{n+1})=\frac{P(s_{t_i}|z^n)P^S(y_{n+1}|x_{n+1},z^n,s_{t_i})}{\sum_k P(s_{t_k}|z^n)P^S(y_{n+1}|x_{n+1},z^n,s_{t_k})}.\tag{35}$$

また, (19) 式より (34) 式右辺は,

$$\begin{aligned}
&q(s_{t_i}|z^{n+1})\prod_{j>i}(1-q(s_{t_j}|z^{n+1}))\\
&\quad=\{q(s_{t_i}|z^n)P^S(y_{n+1}|x_{n+1},z^n,s_{t_i})\\
&\quad\quad/(q(s_{t_i}|z^n)P^S(y_{n+1}|x_{n+1},z^n,s_{t_i})\\
&\quad\quad+(1-q(s_{t_i}|z^n))q(y_{n+1}|x_{n+1},z^n,s_{t_{i-1}}))\}\\
&\quad\quad\times\textstyle\prod_{j>i}\{1\\
&\quad\quad-\{q(s_{t_j}|z^n)P^S(y_{n+1}|x_{n+1},z^n,s_{t_j})\\
&\quad\quad/(q(s_{t_j}|z^n)P^S(y_{n+1}|x_{n+1},z^n,s_{t_j})\\
&\quad\quad+(1-q(s_{t_j}|z^n))q(y_{n+1}|x_{n+1},z^n,s_{t_{j-1}}))\}\}\\
&\quad=\frac{q(s_{t_i}|z^n)P^S(y_{n+1}|x_{n+1},z^n,s_{t_i})}{q(y_{n+1}|x_{n+1},z^n,s_{t_i})}\\
&\quad\quad\times\textstyle\prod_{j>i}\{\frac{(1-q(s_{t_j}|z^n))q(y_{n+1}|x_{n+1},z^n,s_{t_{j-1}})}{q(y_{n+1}|x_{n+1},z^n,s_{t_j})}\}\\
&\quad=q(s_{t_i}|z^n)\textstyle\prod_{j>i}(1-q(s_{t_j}|z^n))\\
&\quad\quad\times\frac{P^S(y_{n+1}|x_{n+1},z^n,s_{t_i})}{q(y_{n+1}|x_{n+1},z^n,s_{t_h})}.
\end{aligned}\tag{36}$$

(16), (27) 式より,

$$\begin{aligned}
&q(s_{t_i}|z^{n+1})\prod_{j>i}(1-q(s_{t_j}|z^{n+1}))\\
&\quad=\frac{P(s_{t_i}|z^n)P^S(y_{n+1}|x_{n+1},z^n,s_{t_i})}{q(y_{n+1}|x_{n+1},z^n,s_{t_h})}\\
&\quad=\frac{P(s_{t_i}|z^n)P^S(y_{n+1}|x_{n+1},z^n,s_{t_i})}{\sum_k P(s_{t_k}|z^n)P^S(y_{n+1}|x_{n+1},z^n,s_{t_k})}.
\end{aligned}\tag{37}$$

よって, (35), (37) 式より, (34) 式が示された.

以上より, 効率的アルゴリズムによる予測がベイズ最適になっていることが証明された.

□

The 25th Symposium on Information Theory
and Its Applications (SITA2002)
Ikaho, Gunma, Japan, Dec. 10–13, 2002

On the Channel Capacity of Universal Channel Coding

Ryo NOMURA * Toshiyasu MATSUSHIMA Shigeichi HIRASAWA

Abstract—In this paper, we introduce the channel capacity in the case that we do not know the probabilistic model of the channel. Then we show the channel capcity in the universal case. To show our main theorem we introduce the decoding scheme that minimizes the probability of error with respect to Bayes criteria.

Keywords—universal channel coding, posterior probability

1 Introduction

In this paper we consider the universal channel coding. In the case that we know the probabilistic structure of the channel, the channel capacity, which is the maximum number of the codeword under the condition that the probability of decoding error vanishes, coincides with the maximum of the mutual information. On the other hand Csiszár et al. showed that in the universal case, the channel capacity also coincides with the maximum of the mutual information by using the type scheme.

We consider the parametric channel. Assuming that we do not know its parameter and we know the parameter class, we introduce the channel capacity in the universal case. Then we show the code that the probability of decoding error is minimum with respect to the Bayes criteria. Moreover we evaluate universal channel capacities by using above decoding scheme.

2 Preliminaries

2.1 Channel capacity

We define a discrete channel to be a system consisting of an finite input alphabet $x \in \mathcal{X}$ and finite output alphabet $y \in \mathcal{Y}$ and a probability transition matrix $P(y|x,\theta)$ that expresses the probability of observing the output symbol y given that we send the symol x where $\theta \in \Theta$ denotes the parameter of the channel which establish the transition probability. Let x^n and y^n denote the sequence of input symbol and that of output symbol, whose length are n.

Definition 2.1 *An (n, M_n) code for the channel consists of the following:*

1. A message W, drawn from the index set $\mathcal{M}_n = \{1, 2, \cdots, M_n\}$.

2. An encoding function $\phi_n : \mathcal{M}_n \to \mathcal{X}^n$, yielding codewords $X^n(1), X^n(2), \cdots, X^n(M_n)$.

3. A decoding function $\varphi_n : \mathcal{Y}^n \to \mathcal{M}_n$.

Then the (average) probability of decoding error is defined as follows.

Definition 2.2 *The (average) probability of error ϵ_n^θ is defined by*

$$\epsilon_n^\theta = \sum_{i=1}^{M_n} P(i) \Pr\{\varphi_n(\phi_n(i) \neq i|\theta\}. \tag{1}$$

Definition 2.3 *The rate r_n of an (n, M_n) code is defined by*

$$r_n = \frac{1}{n} \log M_n.$$

We define the achievability of the rate as follows.

Definition 2.4 *R is called achievable rate when there exists (n, M_n) code such that*

$$\lim_{n\to\infty} \epsilon_n^\theta = 0 \tag{2}$$

$$\lim_{n\to\infty} \frac{1}{n} \log M_n > R. \tag{3}$$

Moreover we shall define the maximum value of achievable rate.

Definition 2.5 *(Channel Capacity)*

$$C(p) = \sup\{R|R \text{ is achievable rate.}\}. \tag{4}$$

Then the case that the true parameter θ^* is known, the following theorem is holding .

Theorem 2.1 *[3] For any stationary memoryless channel we have*

$$C(p) = \max_{P(x)} I(X;Y|\theta^*), \tag{5}$$

where

$$I(X;Y|\theta^*) = \frac{1}{n} \sum_{y^n} \sum_{x^n} P(x^n, y^n|\theta^*) \log \frac{P(x^n|y^n, \theta^*)}{P(x^n)},$$

denotes the mutual information.

Above theorem shows that the maximum codelength per symbol is upper bounded by the maximum mutual information between the input probability distribution $P(x)$ and $P(x^n|y^n, \theta^*)$, under the condition that the probability of error converges to 0.

2.2 Previous Work

In the case that θ^* is unknown, since the probability of error depends upon the probability $P(x|y, \theta^*)$, the achievable rate is defined in several manners.

Definition 2.6 *(Universal achievable rate)*
R is called universal achievable rate when there exists (n, M_n) code such that

$$\lim_{n\to\infty} \epsilon_n^{\theta^*} = 0 \tag{6}$$

$$\lim_{n\to\infty} \frac{1}{n} \log M_n > R. \tag{7}$$

Moreover we shall define the maximum value of universal achievable rate.

* School of Science and Engineering, WASEDA University.
Email: nomu@matsu.mgmt.waseda.ac.jp

Definition 2.7 *(Universal Channel Capacity)*

$$C_U(\theta^*) = \sup\{R|R \text{ is universal achievable rate }\}. \tag{8}$$

Then the following lemma was shown.

Lemma 2.1 *[2]*

$$C_U(\theta^*) = \max_{P(x)} I(X;Y|\theta^*), \tag{9}$$

2.3 Universal Channel Capacity

In the case that the probability distribution $P(\theta)$ is known, the universal channel capacity with respect to $P(\theta)$ can be considered.

Definition 2.8 *(Universal achievable rate w.r.t. $P(\theta)$) R is called universal achievable rate with respect to $P(\theta)$ when there exists (n, M_n) code such that*

$$\lim_{n\to\infty} \int_\theta P(\theta)\epsilon_n^\theta d\theta = 0 \tag{10}$$

$$\lim_{n\to\infty} \frac{1}{n} \log M_n > R. \tag{11}$$

Moreover we shall define the maximum value of universal achievable rate with respect to $P(\theta)$.

Definition 2.9 *(Universal Channel Capacity w.r.t. $P(\theta)$)*

$$C_U(P(\theta)) = \sup\{R|R \text{ is universal achievable rate w.r.t. } P(\theta)\}. \tag{12}$$

3 Minimum error probability decoding scheme

In this section we consider the decoding scheme that minimizes the probability of error with respect to Bayes criteria to evaluate the universal channel capacity and Sup-universal channel capacity. We introduce the following decoding scheme $\varphi_B(\cdot)$.

$$\varphi_B(y^n) = \arg\max_i \int_{\theta\in\Theta} P(X^n(i)|y^n,\theta)P(\theta|y^n)d\theta. \tag{13}$$

where $P(X^n(i)|y^n,\theta)$ is calculated by using Bayes theorem such as

$$P(X^n(i)|y^n,\theta) = \frac{P(y^n|X^n(i),\theta)P(X^n(i)|\theta)}{P(y^n|\theta)}.$$

Then we have the following lemma.

Lemma 3.1 *The decoding function $\varphi_B(\cdot)$ minimizes the probability of error with respect to Bayes criteria.*

(proof)

Assuming that the loss function $L(X^n(i),\varphi(y^n))$ is as follows.

$$L(X^n(i),\varphi(y^n)) = \begin{cases} 0 & i = \varphi(y^n) \\ 1 & i \neq \varphi(y^n) \end{cases}$$

Then the risk function is equal to the probability of error. The Bayes risk is denoted by

$$\int_{\theta\in\Theta} \sum_i \sum_{y^n} L(X^n(i),\varphi(y^n))P(y^n|X^n(i),\theta) \cdot P(X^n(i))P(\theta)d(\theta). \tag{14}$$

Given y^n, to minimize the Bayes risk is equivalent to minimize the following quantity[5].

$$\int_{\theta\in\Theta} \sum_i L(X^n(i),\varphi(y^n))P(X^n(i)|y^n,\theta) \cdot P(\theta|y^n)d(\theta) = \int_{\theta\in\Theta} \sum_{X^n(i)\in C_n} P(X^n(i)|y^n,\theta)P(\theta|y^n)d\theta, \tag{15}$$

where $C_n = \{X^n(i)|i \neq \varphi(\phi(X^n(i)))\}$.

So to choose sequence $X^n(i)$ such that

$$\arg\max_i \int_{\theta\in\Theta} P(X^n(i)|y^n,\theta)P(\theta|y^n)d\theta,$$

then the Bayes risk is minimized. Therefore we deduce the lemma.

We call above decoding function the minimum error probability decoding function in this paper.

4 Evaluation of the Universal Channel Capacity

In this section we show the universal channel capacity w.r.t. $P(\theta)$, by considering the minimum error probability decoding function.

4.1 Universal Channel Capacity

Our main theorem is as follows.

Theorem 4.1 *For the parametric channel that has a prior distribution $P(\theta)$, we have*

$$C_U(P(\theta)) = \max_{P(x)} \left\{ \int_\theta P(\theta) \sum_{x^n,y^n} P(x^n,y^n|\theta) \log \frac{AP(x^n|y^n)}{P(x^n)} d\theta \right\},$$

where

$$AP(x^n|y^n) = \int_\theta P(\theta)P(x^n|y^n,\theta)d\theta,$$

defined by previous section.

(proof)

(Direct Part) We use the random coding technique and the minimum error probability decoding. Assuming that $X^n(i) \sim P(x)$, where $P(x)$ is arbitrary probabilitic distribution and a message $i_0 \in M_n$ was sent. At first assuming that the parameter of the channel θ is fixed.

From the definition of the probability of the error and the minimum error probability decoding scheme, the probability of error given i_0 for some fixed other codeword $x^n(i)$ is given by

$$\epsilon_n^\theta(i_0,i) = \sum_{y^n\in\mathcal{Y}^n} \Pr\{-\log AP(x^n(i_0)|y^n) \geq -\log AP(x^n(i)|y^n)\}.$$

Noting that $x^n(i)$ is independent of the sequence y^n we have

$$\Pr\{-\log \frac{AP(x^n(i_0)|y^n)}{AP(x^n(i)|y^n)} \geq 0\} = \Pr\{-\log \frac{AP(x^n(i_0)|y^n)}{P(x^n(i))} \geq 0\}. \tag{16}$$

So we have

$$\begin{aligned}
&\epsilon_n^\theta(i_0,i)\\
&= \sum_{y^n\in\mathcal{Y}^n} \Pr\{-\log\frac{AP(x^n(i_0)|y^n)}{P(x^n(i))}\geq 0\}\\
&= \sum_{(x^n(i_0),y^n)\in T_n}\ \sum_{x^n(i),i\neq i_0} P(x^n(i_0),y^n)P(x^n(i))\\
&\leq \sum_{(x^n(i_0),y^n)\in T_n} P(x^n(i_0),y^n), \qquad (17)
\end{aligned}$$

where $T(x^n,y^n)=\{(x^n,y^n)|-\log\frac{AP(x^n|y^n)}{P(x^n)}\geq 0\}$.

Let A_n be as follows,

$$\begin{aligned}
A_n &= \left\{(x^n,y^n)\left|\left|\frac{1}{n}\log\frac{AP(x^n|y^n)}{P(x^n)}\right.\right.\right.\\
&\quad \left.\left.\left.-E_xE_y\log\frac{AP(X^n|Y^n)}{P(X^n)}\right|<\gamma\right\}, \qquad (18)
\end{aligned}$$

for some fixed $\gamma>0$ From the law of large numbers we have

$$\Pr\{A_n\}=\sum_{(x^n,y^n)\in A_n} P(x^n,y^n)\to 1. \qquad (19)$$

So for sufficient large n we have

$$\begin{aligned}
&\sum_{(x^n(i_0),y^n)\in T_n} P(x^n(i_0),y^n)\\
&= \sum_{(x^n(i_0),y^n)\in A_n\cap T_n} P(x^n(i_0),y^n)\\
&\quad + \sum_{(x^n(i_0),y^n)\in \bar{A}_n\cap T_n} P(x^n(i_0),y^n)\\
&\leq \sum_{(x^n(i_0),y^n)\in A_n\cap T_n} P(x^n(i_0),y^n)+\eta\\
&= \sum_{(x^n(i_0),y^n)\in A_n\cap T_n} P(x^n(i_0))P(y^n|x^n(i_0))\\
&\quad +\eta, \qquad (20)
\end{aligned}$$

where $\bar{A}_n$ denotes the complement of A_n and $\eta>0$.

On the other hand from the definition of A_n for any $(x^n,y^n)\in A_n$, we have

$$\begin{aligned}
&P(x^n)\\
&< AP(x^n|y^n)\exp\left\{-n\left(E_xE_y\log\frac{AP(x^n|y^n)}{P(x^n)}-\gamma\right)\right\}\\
&\leq \exp\left\{-n\left(E_xE_y\log\frac{AP(x^n|y^n)}{P(x^n)}-\gamma\right)\right\} \qquad (21)
\end{aligned}$$

Substituting (21) into (20), for sufficient large n we have

$$\begin{aligned}
&\sum_{(x^n(i_0),y^n)\in T_n} P(x^n(i_0),y^n)\\
&= \sum_{(x^n(i_0),y^n)\in A_n\cap T_n} P(x^n(i_0))P(y^n|x^n(i_0))+\eta\\
&\leq \sum_{(x^n(i_0),y^n)\in A_n\cap T_n} P(y^n|x^n(i_0))\\
&\quad \cdot\exp\left\{-n\left(E_xE_y\log\frac{AP(x^n|y^n)}{P(x^n)}-\gamma\right)\right\}+\eta\\
&\leq \exp\left\{-n\left(E_xE_y\log\frac{AP(x^n|y^n)}{P(x^n)}-\gamma\right)\right\}\\
&\quad +\eta \qquad (22)
\end{aligned}$$

Let $\epsilon_n^\theta(i_0)$ denotes the probability of error given i_0 and θ. By using the union bound and from (17) and (22) for sufficient large n we have

$$\begin{aligned}
\epsilon_n^\theta(i_0) &\leq \sum_{i\neq i_0, 1\leq i\leq M_n}\epsilon_n^{\theta^*}(i_0,i) \qquad (23)\\
&\leq M_n\\
&\quad \cdot\exp\left\{-n\left(E_xE_y\log\frac{AP(x^n|y^n)}{P(x^n)}-\gamma\right)\right\}\\
&\quad +M_n\eta
\end{aligned}$$

Because of using the random coding procedure $\epsilon_n^\theta(i_0)$ is not depend on $i_0\in M_n$. So we have

$$\begin{aligned}
\int_\theta P(\theta)\epsilon_n^\theta d\theta &= \int_\theta P(\theta)\frac{1}{M_n}\sum_{i_0=1}^{M_n}\epsilon_n^\theta(i_0)d\theta\\
&= \int_\theta P(\theta)\epsilon_n^\theta(i_0)d\theta. \qquad (24)
\end{aligned}$$

Hence for sufficient large n, the error probability is denoted by

$$\begin{aligned}
&\int_\theta P(\theta)\epsilon_n^\theta d\theta \leq \int_\theta P(\theta)M_n \qquad (25)\\
&\cdot\left(\exp\left\{-n\left(E_xE_y\log\frac{AP(x^n|y^n)}{P(x^n)}-\gamma\right)\right\}+\eta\right)d\theta.
\end{aligned}$$

So from the concavity, assuming that $\frac{1}{n}\log M_n > \int_\theta P(\theta)E_xE_y\log\frac{AP(x^n|y^n)}{P(x^n)}d\theta$, we can show that

$$\int_\theta P(\theta)\epsilon_n^\theta d\theta\to 0.$$

Note that we use the random coding, there exists a (n,M_n) code that satisfies $\lim_{n\to\infty}\int_\theta P(\theta)\epsilon_n^\theta d\theta=0$.

Therefore Direct part is deduced.

(Converse Part) From the similar argument to [1], we have

$$\begin{aligned}
&\int_\theta P(\theta)\epsilon_n^\theta d\theta\\
&\geq \Pr\left\{\frac{1}{n}\log\frac{AP(X^n|Y^n)}{P(X^n)}\leq\frac{1}{n}\log M_n-\gamma\right\}\\
&\quad -e^{-n\gamma}, \qquad (26)
\end{aligned}$$

where $\gamma>0$. Assuming that rate

$$\begin{aligned}
&\frac{1}{n}\log M_n\geq R\\
&= \max_{P(x)}\left\{\int_\theta P(\theta)\sum_{x^n,y^n}P(x^n,y^n|\theta)\log\frac{AP(x^n|y^n)}{P(x^n)}d\theta\right\}\\
&\quad +2\gamma \qquad (27)
\end{aligned}$$

is achievable, and we shall lead a contradiction.

Substituting (27) into (26), we have

$$\begin{aligned}\int_\theta P(\theta)\epsilon_n^\theta d\theta \\ \geq \Pr\left\{\frac{1}{n}\log\frac{AP(X^n|Y^n)}{P(X^n)} \leq \frac{1}{n}\log M_n - \gamma\right\} \\ -e^{-n\gamma} \\ > 0. \end{aligned} \tag{28}$$

So we lead a contradiction. Therefore for any

$$R \leq \max_{P(x)}\left\{\int_\theta P(\theta)\sum_{x^n,y^n} P(x^n,y^n|\theta)\log\frac{AP(x^n|y^n)}{P(x^n)}d\theta\right\}, \tag{29}$$

is not achievable.

5 Consideration

We consider the probability of error for some fixed θ by using the minimum error probability decoding scheme. From the same argument of Theorem 4.1 following theorem can be shown.

Definition 5.1

R is called universal achievable rate by using the minimum error probability decoding scheme when there exists (n, M_n) code such that

$$\lim_{n\to\infty}\epsilon_n^{\theta^*} = 0 \tag{30}$$

$$\lim_{n\to\infty}\frac{1}{n}\log M_n > R, \tag{31}$$

by using the minimum decoding scheme.

Moreover we shall define the maximum value of universal achievable rate by using the minimum error probability decoding scheme.

Definition 5.2

$$\begin{aligned}C_U^m(\theta) = \sup\{R|&R \textit{ is universal achievable rate} \\ &\textit{by using the minimum error probability} \\ &\textit{decoding scheme.}\}.\end{aligned} \tag{32}$$

Theorem 5.1 *For the parametric channel and some fixed θ^**

$$\begin{aligned}C_U^m(\theta) = \max_{P(x)}\{&I(X;Y|\theta^*) \\ &-D(P(X^n|\theta^*,Y^n);AP(X^n|Y^n))\}\end{aligned} \tag{33}$$

where

$$\begin{aligned}&I(X;Y|\theta^*) \\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{y^n}\sum_{x^n} P(x^n,y^n|\theta^*)\log\frac{P(x^n|y^n,\theta^*)}{P(x^n)},\end{aligned} \tag{34}$$

denotes the mutual information rate between $P(x^n)$ and $P(y^n)$ given θ^* and

$$\begin{aligned}&D(P(X^n|\theta^*,Y^n);AP(X^n|Y^n)) \\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{y^n}\sum_{x^n} P(x^n,y^n|\theta^*) \\ &\quad\cdot\log\frac{\int_\theta P(x^n|y^n,\theta)P(\theta|y^n)d\theta}{P(x^n|y^n,\theta^*)},\end{aligned} \tag{35}$$

denotes the KL-divergence between $P(x^n|y^n,\theta^*)$ and $\int_\theta P(x^n|y^n,\theta)P(\theta|y^n)d\theta$.

Above theorem shows that the minimum error probability decoding scheme is asymptotically optimal.

6 Conclusion

In this paper, we defined the universa channel capcity with respect to the prior distribution $P(\theta)$. Then first we have shown the decoding scheme that minimizes the probability of error with respect to Bayes criteria. However the calculation of that is hard because of computational complexity. So to consider an efficient algorithm is future work. Secondly we evaluate the universal channel capacity with respect to $P(\theta)$. In the previous research the difference between the channel capacity and the universal channel capacity is not cleared. From our results the relation between the channel capacity and the universal channel capacity has been revealed.

Acknowledgements

The authors would like to acknowledge all of the member of Hirasawa Lab. and Matsushima Lab. for their helpful suggestions to this work. This research was supported in part by the Ministry of Education under Grant-Aids 12050400 for Scientific Research and Waseda University under Grant 2001A-570 and 2001A-594 for Special Research Projects.

References

[1] T. S. Han, "Information-Spectrum Methods in Information Theory," Baifukan-Press, Tokyo, 1998(In Japanese).

[2] Csiszár and Körner. "Information Theory, Coding Theorems for Discrete Memoryless Systems," Academic Press, 1981.

[3] T.M.Cover and J.A.Thomas, "Elements of information theory," Wiley,1991.

[4] P.Billingsley, "Probability and Measure," Wiley, 1995.

[5] J. M. Bernardo, A.F.M.Smith, "Bayesian Theory," New York: Wiley, 1994.

[6] J.Wolfowitz, "Coding Theorems of Information Theory," Springer-Verlag, NewYork, 1968.

[7] R.G.Gallager, "Information Theory and Reliable Communication", Wiley, 1968.

[8] J.Ziv, "Universal decoding for finite-state channels," IEEE Trans. Inf. Theory, vol.31,no.4, pp.453-460, July 1985.

[9] M.Feder and A.Lapidoth, "Universal decoding for channels with memory," IEEE Trans. Inf. Theory, vol.44,no.5, pp.1726-1745, Sept. 1998.

[10] T.Uyematsu, "On the Universality of Channel Decoders Constructed from Source Encoders for Finite-State Channels," IEICE Trans. Fundamentals, Vol.E84-A No.10 pp.2436-2446

The 25th Symposium on Information Theory
and Its Applications (SITA2002)
Ikaho, Gunma, Japan, Dec. 10–13, 2002

誤り訂正符号構成法を利用した直交計画の構成法に関する一考察
A Note on the Construction of Orthogonal Designs by using the Construction of Error Correcting Codes

斉藤友彦* 吉田隆弘* 松嶋敏泰*
Tomohiko Saito Takahiro Yoshida Toshiyasu Matsushima

Abstract— In the field of the experimental design in statistics, a orthgonal design was studied to get the many informataion with the few number of times of an experiment. Orthgonal desing was expressed as linear vector subspace of vector space called complete design. And linear code was expressed as linear vector subspace of vector space. There are many common features in both. In this paper,we clarify both relation, and show that the construction of linear code is useful for the construction of orthgonal design.

Keywords— Experimental design,Orthogonal design,Error-Correcting code

1 はじめに

統計学における実験計画法の分野では，少ない実験回数でより多くの情報を得るために直交計画がよく用いられる [1]．直交計画では，できるだけ少ない実験回数で，与えられた要因 (因子) およびそれらの間にいくつかの想定される交互作用を全て分離推定できるように計画を立てることが重要である．

従来，直交計画は射影幾何を用いたわりつけによって構成されているが，実験計画法では現実問題において3次以上の交互作用は無視して良い場合が多く，高次の交互作用を考慮に入れた直交計画を構成する問題は従来，あまり考えられていない．しかし，今後実験計画法が学習など他の分野へ応用される可能性があることを考えると [2]，高次の交互作用も考慮に入れた理論的考察を行うことは十分意味のあることである．

ところで，直交計画は完全計画と呼ばれる線形ベクトル空間の部分空間として表現される．そして，直交計画を構成する問題は各列が，交互作用に依存する，ある一次独立な関係を満たす生成行列を構成する問題に帰着する．例えば，交互作用がない場合，直交計画を構成する問題は，任意の2列が一次独立な行列を構成する問題に帰着する．また，誤り訂正符号における線形符号も同様に，線形ベクトル空間の部分空間として表現することができ，線形符号構成問題は各列が，最小距離に依存する，ある一次独立な関係を満たす (検査) 行列を構成する問題に帰着する [4]．両者には多くの共通点が存在するが，その関係は従来ほとんど明らかにされてはいない．

そこで，本稿では2つの分野の関係に関する考察を行い，誤り訂正符号の構成法が，高次の交互作用も考慮に入れた場合における，直交計画の構成法に利用できることを示す．本稿では特に以下の2つに関して考察を行う．1つは誤り訂正符号の分野における不均一誤り訂正符号 [3] と直交計画の関係についてである．ここでは，不均一誤り訂正符号を構成する問題が，(均一な) 線形符号構成問題よりも，直交計画を構成する問題により近いことを示す．もう1つは，分割法と呼ばれる実験に関する直交計画 [5] と誤り訂正符号の関係についてである．ここでは，分割法に適した計画を構成することが，ある性能がよい符号を部分符号として持つ符号を構成することに等しいことを示す．

* 〒169-8555 東京都新宿区大久保 3-4-1 早稲田大学経営システム工学科. Dept of Industrial and Management Systems Engineering, Waseda University, Okubo3-4-1,Shinjuku,Tokyo,169-8555 Japan. E-mail:tomohiko@matsu.mgmt.waseda.ac.jp

2 実験計画法

2.1 直交計画

いま n 個の因子を $F_1, F_2, \cdots, F_n$ とし，各因子に与える条件 (水準) の集合を $\{\Omega_1, \Omega_2, \cdots, \Omega_n\}$ とする．ここで，水準の集合の大きさ，すなわち水準数は $|\Omega_i| = \omega_i, (i = 1, 2, \cdots n)$ とする．また，k 因子間に存在する交互作用を k 次の交互作用と呼び，$F_{i_1} \times \cdots \times F_{i_k}$ とかく．また，$M = \{1, 2, \cdots, n\}$, $I = \{\{i_1, i'_1\}, \{i_2, i'_2\}, \cdots, \{i_w, i'_w\}\} \subseteq M^2$ とし，M は因子の番号 (インデックス) の集合，I は2次の交互作用が考えられる二つの因子の番号の対のインデックス集合とする．

ここで，

$$\begin{aligned}\boldsymbol{\Omega} &= \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n \\ &= \{(\nu_1, \nu_2, \cdots, \nu_n); \forall \nu_i \in \Omega_i, i = 1, \cdots, n\},\end{aligned}$$

とする．因子 F_i を ν_i に対応させ，ベクトル $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_n)$ を $\boldsymbol{\Omega}$ 上の点とする．このとき点 $\boldsymbol{\nu}$ の集合 $\boldsymbol{\Gamma}(\subseteq \boldsymbol{\Omega})$ を計画と呼ぶ．また，$\boldsymbol{\Omega}$ を完全計画と呼び，$\boldsymbol{\Omega}$ 上の全ての点に関して，すなわち全ての水準組合せで，実験を行うことをを完全実験と呼ぶ．

定義 (直交計画)

計画 $\boldsymbol{\Gamma}(\subseteq \boldsymbol{\Omega})$ を考える．$\boldsymbol{\Gamma}$ の中で因子 $F_{i_1}, \cdots, F_{i_t}$ の水準が $\varphi_1, \cdots, \varphi_t$ であるような点全体を

$$\boldsymbol{\Gamma}^{i_1, \cdots, i_t}_{\varphi_1, \cdots, \varphi_t} = \{(\nu_1, \cdots, \nu_n) \in \boldsymbol{\Gamma}; \nu_{i_1} = \varphi_1, \cdots, \nu_{i_t} = \varphi_t\}, \quad (\varphi_j \in \Omega_{i_j})$$

とする．このとき，

$$|\boldsymbol{\Gamma}^{i_1, \cdots, i_t}_{\varphi_1, \cdots, \varphi_t}| = \frac{N}{\omega_{i_1} \cdots \omega_{i_t}}, \quad (\forall \varphi_1 \in \Omega_{i_1}, \cdots, \forall \varphi_t \in \Omega_{i_t})$$

となるような計画 $\boldsymbol{\Gamma}$ を強さ t の直交計画と呼び，$\boldsymbol{\Gamma}$ 上の全ての点に関して実験を行うことを直交実験と呼ぶ．ただし，$N = |\boldsymbol{\Gamma}|$ であり，N を計画 $\boldsymbol{\Gamma}$ の大きさとよぶ．□

因子 M，及び2次の交互作用の組 I が与えられた下で，その効果を測るためには，次の条件を満たす直交計画 $\mathbf{\Gamma}$ が必要である．

1. $\mathbf{\Gamma}$ は強さ2でなければならない．
2. $\mathbf{\Gamma}$ は部分的に強さ3，つまり任意の $\{i_1,i_2\}\in I, i_3\notin\{i_1,i_2\}$ に対して

$$|\mathbf{\Gamma}^{i_1,i_2,i_3}_{\varphi_1,\varphi_2,\varphi_3}|=\frac{N}{\omega_{i_1}\omega_{i_2}\omega_{i_3}},(\forall\varphi_j\in\Omega_{i_j},j=1,2,3)$$

でなければならない．

3. $\mathbf{\Gamma}$ は部分的に強さ4，つまり $\{i_1,i_2\}\cap\{i_3,i_4\}=$ "空集合"なる任意の $\{i_1,i_2\},\{i_3,i_4\}\in I$ に対して

$$|\mathbf{\Gamma}^{i_1,i_2,i_3,i_4}_{\varphi_1,\varphi_2,\varphi_3,\varphi_4}|=\frac{N}{\omega_i\omega_{i_2}\omega_{i_3}\omega_{i_4}},(\varphi_j\in\Omega_{i_j},j=1,\cdots,4)$$

でなければならない．

ここで，以下，ω_i を一定，すなわち $\omega_i=q,\Omega_i=GF(q),(i=1,\cdots,n)$ とする．ただし，q は素数の累乗である．このとき，直交計画 $\mathbf{\Gamma}$ は生成行列 G の行ベクトルによって張られる $GF(q)^n$ 空間の部分空間

$$\mathbf{\Gamma}=\{\boldsymbol{\nu}=\boldsymbol{\theta}G;\boldsymbol{\theta}\in GF(q)^m\},\qquad(1)$$

として表現される．ただし，

$$G=\begin{bmatrix}g_{11}&\cdots&g_{1n}\\ \vdots&&\vdots\\ g_{m1}&\cdots&g_{mn}\end{bmatrix}=\begin{bmatrix}\boldsymbol{g_1}&\boldsymbol{g_2}&\cdots&\boldsymbol{g_n}\end{bmatrix},\qquad(2)$$

である．このとき上記三つの性質は次のように変わる．

1'. G の任意の2列は互いに一次独立である．
2'. 任意の $\{i_1,i_2\}\in I, i_3\notin\{i_1,i_2\}$ に対して，$\boldsymbol{g_{i_1}},\boldsymbol{g_{i_2}},\boldsymbol{g_{i_3}}$ は $GF(q)$ 上一次独立
3'. $\{i_1,i_2\}\cap\{i_3,i_4\}=$ "空集合"なる任意の $\{i_1,i_2\},\{i_3,i_4\}\in I$ に対して，$\boldsymbol{g_{i_1}},\boldsymbol{g_{i_2}},\boldsymbol{g_{i_3}},\boldsymbol{g_{i_4}}$ は $GF(q)$ 上一次独立

このとき，因子 F_i は $\boldsymbol{g_i}$ に対応しており (つまり ν_i と $\boldsymbol{g_i}$ が対応している)，この対応をわりつけと呼ぶ．

以上より，直交計画を構成する問題は因子の数 n，交互作用の組 I が与えられた下で，実験の大きさ，すなわち (2) 式における m の値が，最も小さくなる (生成) 行列を構成する問題に帰着する．

特別な場合として，2次の交互作用が全てある場合，つまり $I=\{\{i,j\};i,j\in\{1,\cdots,n\},i\neq j\}$ の場合，強さ4の直交計画が必要となる．これはつまり任意の4列が一次独立な生成行列 G が必要となる．一般に，k 次の交互作用を全て考えるときは，任意の $2k$ 列が一次独立な生成行列を構成すればよい．

2.2 有限射影幾何によるわりつけ

大きさ $N=q^m$ の直交計画を考えるときには，$PG(m-1,q)$ なる射影幾何の点に各因子をわりつける．たとえば，因子 F_i を $PG(m-1,q)$ の点 $\boldsymbol{g_i}$ にわりつけた場合 $(F_i\to\boldsymbol{g_i})$，生成行列は

$$G=\begin{bmatrix}\boldsymbol{g_1}&\boldsymbol{g_2}&\cdots&\boldsymbol{g_n}\end{bmatrix},$$

で与えられる．

これは，誤り訂正符号におけるハミング符号の検査行列に因子 F_i をわりつけたものに等しい．ただし，わりつけの際，射影幾何における直線や平面の性質を用いて2次の交互作用に対応している．この方法では，高次の交互作用があるような場合，計算量が多くなり，非常に難しい問題になる．

3 誤り訂正符号

3.1 線形符号

n 次元線形ベクトル空間 $\{0,1,\cdots,q-1\}^n$ の k 次元部分空間を q 元 (n,k) 線形符号と呼ぶ．ここで，n を符号長，k を次元 (情報長) と呼ぶ．(n,k) 線形符号 C は $(n-k)\times n$ の検査行列 H によって次のように表現することができる．

$$C=\{\boldsymbol{c}\in GF(q)^n;\boldsymbol{c}H^T=\boldsymbol{0}\}.$$

このとき，検査行列 H と最小距離 d には次の定理のような関係がある．

定理1 [4]

線形符号 C の最小距離が d である場合を考える．このとき線形符号 C の検査行列 H の任意の $d-1$ 個以下の列ベクトルは $GF(q)$ 上一次独立である． □

線形符号を構成する問題は，符号長 n，最小距離 d が与えられた下で，$n-k$ の値が最も小さくなる (検査) 行列 H を構成する問題に帰着する．

3.2 不均一誤り訂正符号

3.2.1 定義

線形符号 C における，各符号語シンボルのセパレーション s_i を次のように定義する．

$$\begin{aligned}s_i=\ & min\{d[\boldsymbol{c},\boldsymbol{c'}];\boldsymbol{c}=(c_1,\cdots,c_i,\cdots,c_n),\\ &\boldsymbol{c'}=(c'_1,\cdots,c'_i,\cdots,c'_n),\boldsymbol{c},\boldsymbol{c'}\in C,c_i\neq c'_i\}\\ =\ & min\{wt(\boldsymbol{c});c_i\neq 0,\boldsymbol{c}\in C\}.\end{aligned}$$

ただし，$d[\boldsymbol{x},\boldsymbol{y}]$ はハミング距離，$w(\boldsymbol{x})$ はハミング重みとする．n 個の符号語のセパレーションの値を並べたものをセパレーションベクトルとし，$\boldsymbol{S}=(s_1,\cdots,s_n)$ とする．

本稿では，この符号語を対照としたセパレーションベクトルに対して，$i\neq j$ かつ $i,j\in\{1,\cdots,n\}$ のとき，ある $s_i\neq s_j$ が存在するような符号を不均一誤り訂正符号と呼ぶ．また，セパレーションと検査行列には次の定理のような関係がある．

定理2 [3]

(n,k) 線形符号における，i 番目の符号語シンボルのセパレーションの値が s_i であるための必要十分条件は，

検査行列の i 番目の列を含む任意の s_i-1 個以下の列が一次独立であることである. □

不均一誤り訂正符号を構成する問題は符号長 n, セパレーションベクトル $\boldsymbol{S}$ が与えられた下で, $n-k$ の値が最も小さくなる (検査) 行列を構成することである.

3.2.2 構成法

本節では, [3] で提案された不均一誤り訂正符号の構成法を述べる.

α を $GF(2^{2m})$ の原始元とする. このとき $\beta=\alpha^{2^m+1}$ は $GF(2^m)(\subset GF(2^{2m}))$ の原始元となる. ここで次の検査行列によって定義される (n,k) 線形符号 V について考える.

$$H=\begin{bmatrix} 1 & \alpha & \cdots & \alpha^{2^m} & \alpha^{2^m+1} & \alpha^{2^m+2} & \cdots & \alpha^{2^{2m}-2} \\ 1 & 0 & \cdots & 0 & \beta^3 & 0 & \cdots & 0 \end{bmatrix} \tag{3}$$

符号 V の符号長は $n=2^{2m}-1$, 次元は $k=2^{2m}-3m-1$ である. また, 符号 V はハミング符号の部分符号なので最小距離は 3 以上である.

定理 3 [3]

m を奇数の整数とする. このとき (3) 式の検査行列 H によって定義される符号を考える. このとき 2^m-1 個以上の符号語シンボルはセパレーションの値が 5 以上になる. □

4 直交計画と誤り訂正符号の関係

4.1 直交計画と線形符号

直交計画及び線形符号 (不均一誤り訂正符号も含む) を構成する問題は, 列の数 n, 及び, 各列のある一次独立な関係, が与えられた下で, 行の数 m, が最も小さくなる, (2) 式のような行列を構成する問題に帰着する. 特に, 直交計画において全ての k 次の交互作用が存在する場合, 最小距離 $2k+1$ の符号を構成する問題と全く同じになる.

したがって, BCH 符号など誤り訂正符号において研究されている符号構成法が, 直交計画を構成する際, 利用できる.

4.2 直交計画と不均一誤り訂正符号

本節では, 直交計画, 線形符号, 不均一誤り訂正符号それぞれの構成法の包含関係を示し, 不均一誤り訂正符号構成問題が直交計画構成問題により近いことを示す.

まず, 簡単な例を用いてこの 3 つの包含関係を示す。次のような 4 列の行列を考える。

$$G=\begin{bmatrix} \boldsymbol{g_1} & \boldsymbol{g_2} & \boldsymbol{g_3} & \boldsymbol{g_4} \end{bmatrix}. \tag{4}$$

因子 F_1,F_2,F_3,F_4 があり, さらに 2 次の交互作用 $F_1\times F_2$ があるとする. このような場合の直交計画を構成する問題を考える. このとき因子 F_i を (4) 式における $\boldsymbol{g_i}$ にわりつける. このとき, (4) 式の行列は以下の列が一次独立でなければならない.

$$\{\boldsymbol{g_1},\boldsymbol{g_2}\},\{\boldsymbol{g_1},\boldsymbol{g_3}\},\{\boldsymbol{g_1},\boldsymbol{g_4}\},\{\boldsymbol{g_2},\boldsymbol{g_3}\},\{\boldsymbol{g_2},\boldsymbol{g_4}\},$$
$$\{\boldsymbol{g_3},\boldsymbol{g_4}\},\{\boldsymbol{g_1},\boldsymbol{g_2},\boldsymbol{g_3}\},\{\boldsymbol{g_1},\boldsymbol{g_2},\boldsymbol{g_4}\}. \tag{5}$$

また, 上記の因子, 交互作用の条件を満たす行列を線形符号の構成法で構成する場合, 最小距離が 4, すなわち, 任意の 3 列が一次独立な検査行列を構成すればよいので, 以下の列が一次独立である (4) 式のような行列 G を求めればよい.

$$\{\boldsymbol{g_1},\boldsymbol{g_2}\},\{\boldsymbol{g_1},\boldsymbol{g_3}\},\{\boldsymbol{g_1},\boldsymbol{g_4}\},\{\boldsymbol{g_2},\boldsymbol{g_3}\},\{\boldsymbol{g_2},\boldsymbol{g_4}\},$$
$$\{\boldsymbol{g_3},\boldsymbol{g_4}\},\{\boldsymbol{g_1},\boldsymbol{g_2},\boldsymbol{g_3}\},\{\boldsymbol{g_1},\boldsymbol{g_2},\boldsymbol{g_4}\},\{\boldsymbol{g_1},\boldsymbol{g_3},\boldsymbol{g_4}\},$$
$$\{\boldsymbol{g_2},\boldsymbol{g_3},\boldsymbol{g_4}\} \tag{6}$$

同様に上記の因子, 交互作用の条件を満たす行列を不均一誤り訂正符号の構成法で構成する場合, セパレーションベクトルが (4,3,3,3) である符号を構成すればよいので, 以下の列が一次独立である (4) 式のような行列 G を構成すればよい.

$$\{\boldsymbol{g_1},\boldsymbol{g_2}\},\{\boldsymbol{g_1},\boldsymbol{g_3}\},\{\boldsymbol{g_1},\boldsymbol{g_4}\},\{\boldsymbol{g_2},\boldsymbol{g_3}\},\{\boldsymbol{g_2},\boldsymbol{g_4}\},$$
$$\{\boldsymbol{g_3},\boldsymbol{g_4}\},\{\boldsymbol{g_1},\boldsymbol{g_2},\boldsymbol{g_3}\},\{\boldsymbol{g_1},\boldsymbol{g_2},\boldsymbol{g_4}\},\{\boldsymbol{g_1},\boldsymbol{g_3},\boldsymbol{g_4}\} \tag{7}$$

(5),(6),(7) 式を比較したとき, (6) 式には $\{\boldsymbol{g_1},\boldsymbol{g_3},\boldsymbol{g_4}\}$, $\{\boldsymbol{g_2},\boldsymbol{g_3},\boldsymbol{g_4}\}$ が, (7) 式には, $\{\boldsymbol{g_1},\boldsymbol{g_3},\boldsymbol{g_4}\}$ が無駄な一次独立な部分として存在する. したがって, 不均一誤り訂正符号のほうが無駄な一次独立な部分が少なく, 直交計画の大きさが小さくなることがある. したがって, 不均一誤り訂正符号のほうが直交計画の構成問題により近いことが分かる.

また, 3 次の交互作用があるような場合, たとえば因子 F_1,F_2,F_3,F_4,F_5,F_6, 交互作用 $F_1\times F_2\times F_3$, $F_4\times F_5$ が存在するとき, セパレーションベクトルが (6,5,3,5,3,3) である符号を作ればよい. これも先ほどと同様線形符号の構成法を用いて最小距離が 6 の符号を作るよりも, 実験の大きさが小さくなり, より直交計画を構成する問題に近いといえる.

4.3 分割法と誤り訂正符号

4.3.1 分割法

本節では, 分割法と呼ばれる水準変更が困難な因子がある場合の実験に適した直交計画について例を用いて説明する.

因子 F_1,F_2,F_3,F_4,F_5 が存在し, 各因子の水準数は 2 とする $(\Omega_i=\{0,1\},i=1,\cdots,5)$. また, 因子間に交互作用はないものとする. ここで, 因子 F_1,F_2,F_3 は水準変更が困難な因子であり, F_4,F_5 は水準変更が容易な因子とする. したがって, 因子 F_1,F_2,F_3 はなるべく水準変更を行わないですむような計画が必要である.

このとき, 次の二つの行列について考える.

$$G_1=\begin{bmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{8}$$

$$G_2 = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (9)$$

それぞれの行列に関して i 列目に因子 F_i をわりつける．このとき，(8) 式を生成行列とした場合の計画は次のようになる．

$$\{(\nu_1, \cdots, \nu_5) = \boldsymbol{\theta} G_1\} = \{(0,0,0,0,0), (0,0,0,1,1), (0,1,1,1,1), (0,1,1,0,0), (1,0,1,0,1), (1,0,1,1,0), (1,1,0,1,0), (1,1,0,0,1)\}. \quad (10)$$

また，(9) 式を生成行列とした場合は次のようになる．

$$\{(\nu_1, \cdots, \nu_5) = \boldsymbol{\theta} G_2\} = \{(0,0,0,0,0), (0,0,1,1,1), (0,1,0,1,1), (0,1,1,0,0), (1,0,0,0,1), (1,0,1,1,0), (1,1,0,1,0), (1,1,1,0,1)\}. \quad (11)$$

このとき，(10) 式では，因子 F_1, F_2, F_3 の水準のパターンが 4 通りあるのに対して，(11) 式では，8 通り存在する．したがって，因子 F_1, F_2, F_3 が水準変更困難な因子の場合，(10) 式のような計画のほうが適しているといえる．

従来，分割法に適した直交計画を作るため，群と呼ばれるものを考慮に入れてわりつけを行っている．次節以降で，その群をより一般化し，分割法と誤り訂正符号の関係を考える．

4.3.2 分割法と生成行列の rank

ある因子に関して，その水準変更の回数を少なくするためには，その因子をわりつけた列が作る行列の rank を小さくすればよい．

これを 4.3.1 節で述べた例を用いて説明する．(8) 式を生成行列として用いた場合，水準変更困難な因子 F_1, F_2, F_3 をわりつけた列が作る行列は次のようになる．

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad (12)$$

また，(9) 式の場合は次のようになる．

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad (13)$$

このとき，(12) 式の行列の rank は 2 であるのに対し，(13) 式の行列は rank が 3 である．したがって，(8) 式を生成行列として用いた場合，因子 F_1, F_2, F_3 に関して，水準変更が少なくすむのである．

4.3.3 分割法と部分符号

本節では分割法に適した直交計画を構成する問題が，性能がよい符号を部分符号 (短縮符号) として持つ符号を構成する問題と等しいことを示す．ここで，性能がよい符号とは符号長を固定した下で，なるべく情報長が長くなる符号である．

これを 4.3.1 節で述べた例を用いて示す．まず (12) 式を検査行列とする符号，すなわち (8) 式を検査行列とする符号の部分符号を考える．このとき，(12) 式の rank は 2 であるので，部分符号は符号長 3，次元 1 となる．同様に，(13) 式を検査行列とする符号，すなわち (9) 式を検査行列とする符号の部分符号を考える．(13) 式の rank は 3 であるので，部分符号は符号長 3，次元 0，すなわち符号語は 0 ベクトルのみとなる．

以上から，(8) 式を検査行列とする符号は，より性能がよい部分符号をもつ符号だといえる．したがって，分割法に適した直交計画を作る問題は，性能がよい符号を部分符号としてもつ符号を構成する問題に帰着され，今後誤り訂正符号における部分符号に関する研究が，分割法に適した直交計画を構成する問題に役立つ可能性がある．

5 まとめ

本稿では，直交計画を構成する問題を，誤り訂正符号構成問題との共通点という視点から考察を行った．本稿では特に，不均一誤り訂正符号と直交計画の関係，分割法と誤り訂正符号の関係に関して考察を行い，不均一誤り訂正符号構成問題が，(均一な) 線形符号構成問題よりも，直交計画を構成する問題により近いこと，及び，分割法に適した直交計画を構成する問題が性能がよい符号を部分符号としてもつ符号を構成することと同じであることを示した．

今後の課題として，さらに直交計画と誤り訂正符号の関係を見出すことが挙げられる．

謝辞

本研究を行うにあたり，ご助言頂いた浮田善文氏，野村亮氏及び松嶋研究室の皆様に深く感謝いたします．なお，本研究の一部は文部省科学研究費基盤 (C)(No.12650400)，早稲田大学特定課題研究助成費 (2001A-570) の援助による．

参考文献

[1] 高橋磐郎, 組合せ理論とその応用, 岩波全書 316, 東京, 1979.

[2] 浮田善文, 松嶋敏泰, 平澤茂一, "直交計画を用いたブール関数の学習に関する一考察," 電子情報通信学会論文誌 A(投稿中).

[3] I.M.Boyarinov,and G.L.Katsman, "Linear Unequal Error Protection Codes," IEEE Trans. Inform. Theory,vol.IT-27,pp168-175,Mar.1981.

[4] 平澤茂一, 西島利尚, 符号理論入門, 培風館, 東京,1999.

[5] 永田靖, 入門実験計画法, 日科技連出版社, 東京,2000.

International Symposium on Information Theory and Its Applications
Xi'an, PRC, October 7–11, 2002

On the evaluation of the achievable codelength of Fixed-length codes

Ryo NOMURA†, Toshiyasu MATSUSHIMA† and Shigeichi HIRASAWA†

†School of Science and Engineering
WASEDA University
51-15-02, 3-4-1 Ohkubo, Shinjuku-ku, Tokyo, 169-8555, Japan
Phone:+81-3-5286-3301, Fax:+81-3-5286-3301
Email: nomu@matsu.mgmt.waseda.ac.jp

Abstract

A performance of lossless Fixed-length codes is evaluated by a probability of error and a codelength. Recently, Han showed that the infimum value of achievable rate, which is the shortest codelength per symbol under the condition that the error probability of a Fixed-length code tends to 0, coincides with the entropy spectrum-sup for general source. However in previous researches the convergence speed to the infimum value of achievable rate has not been revealed. In this paper, we consider the infimum value of the achievable codelength, which is the minimum description length to distinguish sequences by using Fixed-length code under the condition that the error probability of Fixed-length code tends to 0. Then we evaluate the infimum value of the achievable codelength by using the variance of self-information.

1. Introduction

In this paper, we consider Fixed-length codes. When we consider Fixed-length codes, the infimum value of the achievable rate, which implies the shortest codelength per symbol under the condition that the probability of decoding error goes to 0, is important. The basic result about the property of Fixed-length codes is that for i.i.d. source the infimum value of the achievable rate coincides with the entropy of the source[1]. Recently, Han extended this result for general source[2].

In previous researches, however, the convergence speed to the infimum value of the achievable rate has not been studied. This is because the achievable rate is based upon the codelength per symbol, a precise evaluation of terms not exceed $o(1)$ was not required.

We consider not the achievable rate, but the achievable codelength, which is the description length to distinguish sequences by using Fixed-length code. Then the infimum value of the achievable codelength can be defined in the same way as the infimum value of the achievable rate. By considering the achievable codelength, we can show the convergence speed to the infimum value of the achievable rate.

This work was supported in part by the Ministry of Education under Grant-Aids 12650400 and Waseda University under Grant 2001A-594 for Special Research Projects.

2. Preliminaries

2.1. Previous Research

We assume that the source alphabet is finite. Let A and $x \in A = \{i : 0 \leq i \leq d\}$ denote a source alphabet and a source symbol respectively, where $d < \infty$. Let $X^n = X_1 X_2 \cdots X_n$ denotes the sequence of random variables from the source, whose outcome is $x^n = x_1 x_2 \cdots x_n$. The probability of the sequence x^n is denoted by $p(x^n)$, that is, $p(x^n) = \Pr\{X^n = x^n\}$. Moreover E denotes the expectation by p, that is, $E[-\log p(X^n)] = \sum_{x^n} p(x^n) \log p(x^n)$.

Then Fixed-length codes are characterized by a pair of Fixed-length encoder and decoder. Let

$$\varphi_n : A^n \to \mathcal{M}_n, \quad \phi_n : \mathcal{M}_n \to A^n,$$

be a Fixed-length encoder and a decoder respectively, where

$$\mathcal{M}_n = \{1, 2, \cdots, M_n\},$$

is called the code set of size M_n. The performance of Fixed-length codes is evaluated by a probability of error and a codelength. We consider the ideal codelength for the sake of simplicity in this paper. So the codelength of a Fixed-length code of size M_n is $\log M_n$. The codelength per symbol of a Fixed-length code of size M_n is $\frac{1}{n} \log M_n$. The probability of error of a Fixed-length code is defined as follows[2].

Definition 2.1 *For an encoder φ_n and a decoder ϕ_n, the probability of error is defined by*

$$\epsilon_n = \Pr\{x^n \neq \phi_n(\varphi_n(x^n))\}. \tag{1}$$

□

Then the achievable rate is defined as follows.

Definition 2.2 R *is called an achievable Fixed-length coding rate for the source if there exists a Fixed-length code such that*

$$\lim_{n\to\infty} \epsilon_n = 0, \tag{2}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log M_n \leq R. \tag{3}$$

□

Moreover we define the infimum value of the achievable Fixed-length coding rate as follows.

Definition 2.3

$$R(X) \equiv \inf\{R \,|\, R \text{ is an achievable Fixed-length coding rate}\},$$

We call $R(X)$ the infimum achievable Fixed-length coding rate. □

In this paper we call the $R(X)$ the infimum achievable rate for short since we only consider Fixed-length codes. Then we know the following result about the property of Fixed-length codes.

Theorem 2.1 *[1] For an i.i.d. source, we have*

$$R(X) = H(X),$$

where $H(X)$ denotes the entropy of the source. □

Recently Han extended the above result for *general* source. In previous researches, however, the convergence speed to achievable rate has not been studied. Our main objective is to evaluate the convergence speed. To evaluate the convergence speed a precise evaluation is required. In previous researches since the achievable rate is the codelength per symbol, a precise evaluation of terms not greater than $o(1)$ was not required. So we shall consider the achievable codelength which is an extension of the achievable rate in this paper.

2.2. The moment of self-information

We shall define quantities that have important roles in this paper.

Definition 2.4 *The rth moment of self-information is defined by*

$$M(X)^r \tag{4}$$
$$= \lim_{n\to\infty} E\left[\left\{\frac{1}{n}\left(\log\frac{1}{p(X^n)} - nH(X)\right)\right\}^r\right],$$

where $H(X)$ denotes the entropy rate. Especially $r = 2$, we call that the variance of self-information. □

The moment of self-information is an inherent value of a source and the variance of self-information coincides with the minimal coding variance, defined by Kontyiannis[6].

Note 2.1 *We use only the variance of self-information in this paper. Note that using the third moment of self-information may give us more precise evaluation of the convergence speed of Fixed-length code.* □

2.3. Definition of an achievable codelength

As we have mentioned above, the infimum achievable rate denotes the shortest codelength per symbol under the condition that the probability of error goes to 0. In this section we shall define the achievable codelength as the codelength to distinguish sequences by using a Fixed-length code. Then the infimum achievable Fixed-length codelength can be defined in the same way as the infimum achievable Fixed-length coding rate. The probability of error is similar to the Definition 2.1.

Definition 2.5 *η_n is called the achievable Fixed-length codelength for the source if there exists the Fixed-length code such that*

$$\lim_{n\to\infty} \epsilon_n = 0, \tag{5}$$

$$\limsup_{n\to\infty} (\log M_n - \eta_n) \le 0. \tag{6}$$

□

Moreover we shall define the infimum value of an achievable Fixed-length codelength.

Definition 2.6

$$L^* = \inf\{\eta_n \,|\, \eta_n \text{ is an achievable Fixed-length codelength}\}. \tag{7}$$

Then we call L^* the infimum achievable Fixed-length codelength. □

We call the L^* the infimum achievable codelength for short since we only consider Fixed-length codes.

Note 2.2 *Assuming that R is a function of n, such as $\sqrt{n}$, the achievable codelength is another expression of the achievable rate. Actually, assuming that $\eta_n = nR$ and dividing by n in (6), it coincides with the achievable rate. To evaluate the convergence speed of Fixed-length codes R must be a function of n. Considering the i.i.d. source, the previous result implies that there exists a Fixed-length code whose convergence speed to $H(X)$ is at most $o(1)$. So to evaluate the convergence speed, the evaluation of the infimum acihevable rate is not sufficient. This is why we introduce the achievable codelength.* □

We shall investigate the infimum achievable codelength in the following section.

3. Evaluation of the infimum achievable codelength

3.1. The minimal error probability code

To evaluate the minimum achievable codelength we consider the code that minimizes the probability of error. We introduce the following code.
(Encoding)

step1 Order the sequence as its probability. Let $x^n(1)$ denotes the highest probability sequence and $x^n(2)$ denotes the second highest probability sequence in the set A^n and so on. That is, for $\forall i < j$, $p(x^n(i)) \ge p(x^n(j))$ is holding.

step2 Let B_n denote the high probability set of x^n whose cardinality is M_n, that is $|B_n| = M_n$. That is,

$$B_n = \{x^n(1), x^n(2), \cdots, x^n(M_n)\}.$$

step3 $\varphi_n^*(x^n(i)) = \begin{cases} i, & i \le M_n \\ 1, & i > M_n. \end{cases}$

(Decoding)

$$\phi_n^*(i) = x^n(i),$$

where $i = \{1, 2, \cdots, M_n\}$. Then we have the following lemma.

Lemma 3.1 *The above code minimizes the probability of error under the condition that the size of code set is M_n.* □

We call above code the minimal error probability code in this paper. Moreover we define the minimal error probability ϵ_n^* as follows

$$\epsilon_n^* = \min_{\varphi_n^*} \Pr\{x^n \neq \phi_n^*(\varphi_n^*(x^n))\}, \tag{8}$$

where φ_n^* and ϕ_n^* denote the minimal error probability encoder and decoder respectively.

3.2. The evaluation of the infimum achievable codelength

We show the infimum Fixed-length achievable codelength, by considering the minimal error probability code in this section. Let φ_* and ϕ_* denote the encoder and the decoder of the minimal error probability code respectively. First we assume the following condition.

Assumption 3.1 *The self-information $-\log p(x^n)$ of a source satisfies the asymptotic normality, that is,*

$$\frac{-\log p(X^n) - nH(X)}{\sqrt{nM(X)^2}} \sim N(0,1). \tag{9}$$

Here, $M(X)^2$ denotes the variance of self-information. In other words, a random variable $-\log p(X^n)$ converges in distribution to Z, where Z is normally distributed with mean $nH(X)$ and variance $M(X)^2$. □

Note 3.1 *In the field of Statistics, sufficient conditions to hold the asymptotic normality has been studied[4]. Note that the asymptotic normality of $-\log p(x^n)$ is holding in an important source class such as an i.i.d. source.* □

Then we have the following theorem.

Theorem 3.1 *Under the Assumption 3.1 we have*

$$L^* = nH(X) + O(\sqrt{nM(X)^2}). \tag{10}$$

(proof) See Appendix. □

From above theorem we know that under the condition that the probability of error goes to 0, the codelength of Fixed-length code is lower bounded by $nH(X) + O\left(\sqrt{nM(X)^2}\right)$. This implies that the infimium value of the convergence speed to $nH(X)$ is lower bounded by $O\left(\sqrt{nM(X)^2}\right)$.

4. Conclusion

In this paper, we investigate the infimum value of the codelength for the source whose self-information satisfies asymptotic normality under the condition that the probability of error goes to 0. The variance of the self-information, which is an inherent value of a source, is efficient to evaluate the infimum achievable codelength. We use only the variance of the self-information and variance of the code, although we defined the rth moment of the code. By using the higher order moment, we may evaluate the code more precisely.

In almost all the previous works, the achievable rate has been considered. However in this paper we consider the achievable codelength to evaluate the convergence speed.

References

[1] R. Gallager, "Information Theory and Reliable Communication," Wiley, 1968.

[2] T. S. Han, "Information-Spectrum Methods in Information Theory," Baifukan-Press, Tokyo, 1998(In Japanese).

[3] T.M.Cover and J.A.Thomas, "Elements of information theory," Wiley,1991.

[4] W.Feller, "An Introduction to Probability Theory and Its Applications," Wiley, 1957, 1966, vol. I-II.

[5] P.Billingsley, "Probability and Measure," Wiley, 1995.

[6] I. Kontoyiannis, "Second-Order Noiseless Source Coding Theorems," *IEEE Trans. Inf. Theory*, 43(4) :1339-1341,1997.

Appendix : Proof of Theorem

(Direct Part) We shall show that any η_n that satisfies $\eta_n > nH(X) + O(\sqrt{nM(X)^2})$, is achievable.

At first we define a code as follows. Let a set K_n be

$$K_n = \{x^n \,||-\log p(x^n) - nH(X)| \leq J_n\}, \tag{11}$$

where, $J_n = \eta_n - nH(X)$.

Since the number of sequences in K_n is $|K_n|$, we can write all the sequences $x^n \in K_n$ as $x^n(c_1), x^n(c_2), \cdots, x^n(c_{|K_n|})$. Moreover we write all the sequences $x^n \notin K_n$ as $x^n(c_{|K_n|+1}), \cdots, x^n(c_{|A^n|})$.

The encoding function $\varphi(\cdot)$ is defined by

$$\varphi(x^n(c_i)) = \begin{cases} i, & i \leq |K_n| \\ 1, & i > |K_n|. \end{cases} \tag{12}$$

The decoding function $\phi(\cdot)$ of this code is $\phi(i) = x^n(c_i)$. The codelength of this code is $\log |K_n|$.

First we shall show the probability of error of this code. The probability of error of this code is denoted by

$$\epsilon_n = \sum_{x^n \notin K_n} p(x^n). \tag{13}$$

On the other hand, we have

$$\begin{aligned} \lim_{n\to\infty} \Pr\{K_n\} &= \lim_{n\to\infty} \sum_{x^n \in K_n} p(x^n) \\ &= \lim_{n\to\infty} \sum_{x^n :- \log p(x^n) < \eta_n} p(x^n) \end{aligned} \tag{14}$$

From Assumption 3.1, we can write

$$\lim_{n\to\infty} \sum_{x^n:-\log p(x^n)<\eta_n} p(x^n) \tag{15}$$
$$= \lim_{n\to\infty} \sum_{x^n:\frac{-\log p(x^n)-nH(X)}{\sqrt{nM(X)^2}}<J_n} p(x^n) = 1.$$

The last equality is derived from the property of normal distribution. So we have

$$\lim_{n\to\infty} \epsilon_n = 0. \tag{16}$$

Second we shall show the size of code set $|K_n|$. From the definition of K_n, for any $x^n \in K_n$ we have

$$-\log p(x^n) \le nH(X) + J_n. \tag{17}$$

So we have for $\forall x^n \in K_n$

$$p(x^n) \ge \exp\{-nH(X) - J_n\}. \tag{18}$$

Then we have the following inequality.

$$\begin{aligned} 1 &\ge \sum_{x^n\in K_n} p(x^n) \\ &\ge |K_n| \exp\{-nH(X) - J_n\}. \end{aligned} \tag{19}$$

So we have

$$|K_n| \le \exp\{nH(X) + J_n\}. \tag{20}$$

Hence the following equality is deduced,

$$\log|K_n| \le nH(X) + J_n = \eta_n. \tag{21}$$

Therefore from (16) and (21), any η_n that satisfies $\eta_n > nH(X) + O(\sqrt{nM(X)^2})$, is achievable.

(Converse Part) We shall show that any η_n that satisfies $\eta_n \le nH(X) + O(\sqrt{nM(X)^2})$is not achievable by using the minimal error probability code. That is we will show if

$$\lim_{n\to\infty} \epsilon_n = 0, \tag{22}$$

is satisfied, then the following inequality is holding

$$\begin{aligned} &\lim_{n\to\infty} \log M_n \\ &> \lim_{n\to\infty} \left(nH(X) + O(\sqrt{nM(X)^2})\right), \end{aligned} \tag{23}$$

where M_n denotes the size of code set of minimal error probability code.

At first, we define the set as follows

$$\begin{aligned} &Q_n(C) \\ &= \left\{x^n \left| -\log p(x^n) - nH(X) > C\sqrt{nM(X)^2}\right.\right\}. \end{aligned} \tag{24}$$

where C is arbitrary constant, such that C. Moreover we define a constant $D > C$ and following set,

$$\begin{aligned} &Q_n(D) \\ &= \left\{x^n \left| -\log p(x^n) - nH(X) \le D\sqrt{nM(X)^2}\right.\right\}. \end{aligned} \tag{25}$$

From Assumption 3.1, for sufficient large n and C there exists constant δ_C which satisfies

$$\Pr\{Q_n(C) \cap Q_n(D)\} \ge \delta_C > 0, \tag{26}$$

where δ_C is a function of C, D. On the other hand, from the definition of $Q_n(C)$, for any sequence $x^n \in Q_n(C)$ we have

$$-\log p(x^n) - nH(X) > C\sqrt{nM(X)^2}. \tag{27}$$

So for any sequence $x^n \in Q_n(C)$, we have

$$p(x^n) \le \exp\left\{-nH(X) - C\sqrt{nM(X)^2}\right\}. \tag{28}$$

From above discussion for sufficient large n we have

$$\begin{aligned} &|Q_n(C) \cap Q_n(D)| \\ &\ge \frac{\Pr\{Q_n(C) \cap Q_n(D)\}}{\max_{x^n\in Q_n(C)\cap Q_n(D)} p(x^n)} \\ &= \delta_C \exp\left\{nH(X) + C\sqrt{nM(X)^2}\right\} \end{aligned} \tag{29}$$

So we have

$$\begin{aligned} &\log|Q_n(C) \cap Q_n(D)| \\ &\ge nH(X) + C\sqrt{nM(X)^2} - O(1). \end{aligned} \tag{30}$$

On the other hand from the definition of $Q_n(D)$, for any sequence $x^n \in Q_n(D)$ we have

$$p(x^n) \ge \exp\{-nH(X) - D\sqrt{nM(X)^2}\}. \tag{31}$$

Note that from (26) and (31), to satisfy (22), it is required that all the sequence $x^n \in Q_n(C) \cap Q_n(D)$ is uniquely decodable. So it is required that

$$M_n \ge |Q_n(C) \cap Q_n(D)|. \tag{32}$$

Substituting (32) into (30), to hold (22) we have

$$\begin{aligned} &\lim_{n\to\infty} \log M_n \\ &> \lim_{n\to\infty} \left(nH(X) + C\sqrt{nM(X)^2} - O(1)\right) \\ &= \lim_{n\to\infty} \left(nH(X) + O(\sqrt{nM(X)^2})\right). \end{aligned} \tag{33}$$

Since we use the minimal error probability code, the probability of error of any code is greater than or equal to ϵ_n^* under the condition that the codelength is the same. Therefore from (33), for any η_n that satisfies $\eta_n \le nH(X) + C\sqrt{nM(X)^2}$ is not acihevable. □

The 25th Symposium on Information Theory
and Its Applications (SITA2002)
Ikaho, Gunma, Japan, Dec. 10–13, 2002

多端子情報源符号化に基づいた分散協調問題の定式化
Distributed cooperative problem based on multiterminal source coding

吉田 隆弘* 松嶋 敏泰* 平澤 茂一*
Takahiro YOSHIDA Toshiyasu MATSUSHIMA Shigeichi HIRASAWA

Abstract— We consider modeling the distributed cooperative systems.In this paper, we apply multi-terminal systems to a distributed cooperative system and formulate a distributed cooperative problem based on decision theory. In the case that loss function is the logarithmic loss function and squared loss function, we shall show the optimum rule minimizing expected loss function and clarify the relationship bitween loss of determination of agents and cost of information exchange of agents.

Keywords— distributed cooperative problem, decision theory, source coding, side information

1 はじめに

コンピュータネットワーク上での分散協調問題とは，ネットワーク上で結合された複数の意思決定者 (エージェント) が個々に観測した情報，およびエージェント間で情報のやりとりをすることによって得られる情報を利用し，全体のシステムとして良い出力となるように，各エージェントが個別に意思決定を行う問題である．本稿で扱う分散協調は，(1) 相関のある複数の情報を各エージェントがそれぞれ個別の情報として観測し，(2) エージェント間で情報交換を行い，(3) 得られた情報から各エージェントが個別に意思決定を行うものとする．このような分散協調の問題設定では，全ての情報を各エージェントが共有してしまえば，一人のエージェントによる基本的な決定問題と同等になってしまうので，エージェント間で交換される情報量に応じたコストと，システム全体としての出力（意思決定）の良さの間にはトレードオフの関係があると考えられる．

一方，情報理論の分野において，複数の送・受信者の情報のやり取りを扱う多端子情報理論がある [1]．この分野における研究の一部では，情報源から発生する相関のある複数の情報を，複数の受信者が個別に受信した情報とは別に，補助的な情報や他の受信者から受け取った部分的な情報などを利用できるものとし，それらの情報を用いて復号を行うといった問題が扱われている [1]．また，多端子情報源符号化を情報源の確率分布パラメータを推定する問題や仮説検定に適用した研究もなされている [3][4][5]．

これらの問題設定における多端子情報源符号化モデルを，本稿で扱う分散協調問題と対応させてみると，各エージェントの意思決定は上述のモデルでの復号，パラメータ推定，および仮説検定を行うこととしてみることができ，各エージェントが観測する個別の情報とエージェント間の情報交換によって得られた情報は符号化された情報としてみることができる．

そこで本稿では，分散協調問題のモデルを多端子情報源符号化モデルとして捉え，エージェント間で情報交換をして，意思決定するという分散協調問題を決定理論 [2] に基づいてモデル化を行う．さらに損失として対数損失と二乗誤差損失を用いた場合の最適な意思決定，および各エージェント間で交換する情報の与えかたを示し，意思決定による損失と情報交換によるコストのトレードオフ関係を明確にする．

* 早稲田大学理工学部経営システム工学科，〒169-8555 新宿区大久保 3-4-1, Dep. of Industrial and Management Systems Engineering, Waseda University, 3-4-1 Ohkubo, Shinjuku-ku, Tokyo 169-8555, E-mail:takahiro@matsu.mgmt.waseda.ac.jp

2 分散協調問題の定式化

2.1 モデル

分散協調問題の基本モデルとして図 1 のような m 人のエージェントの協調問題を考える．時刻 t での情報源からの出力を確率変数ベクトル $X_1^n[t] = (X_1[t], \cdots, X_n[t])$ とし，それぞれの要素 $X_i[t]$ は有限集合 $\mathcal{X}_i$ $(i = 1, \cdots, n)$ のなかに値をとる $(m \leq n)$．さらに，それら有限集合の直積 $\mathcal{X}_i^t$ の中に値をとる確率変数列を

$$X_i[1,t] = X_i[1]X_i[2]\cdots X_i[t] \tag{1}$$

と表記する．また確率変数ベクトル $X_1^n[t]$ の同時確率分布，

$$P\left(x_1^n[t]\right) = \Pr\left\{X_1^n[t] = x_1^n[t]\right\} \tag{2}$$

は全エージェントに対して既知であると仮定する．また，時刻 1 から t までの情報源の出力系列 $X_1^n[1,t] = X_1^n[1]X_1^n[2]\cdots X_1^n[t]$ の同時確率は，

$$P\left(x_1^n[1,t]\right) = \prod_{i=1}^{t} P\left(x_1^n[i] \mid x_1^n[1,i-1]\right) \tag{3}$$

となる．

図 1 分散協調問題のモデル

各エージェント DM_i $(i = 1, \cdots, m)$ は時刻 t において，情報源からの部分的な出力系列 $x_i[1,t]$ と他のエージェントが観測した出力系列を符号化した情報

$$\overline{\mathbf{u}}_i[t] = \{u_1[t], \cdots, u_{i-1}[t], u_{i+1}[t], \cdots u_m[t]\} \tag{4}$$

を受信できるものとする．ここで，$u_i[t] \in \mathcal{U}_i[t]$,

$$\varphi_i^t : \mathcal{X}_i^t \to \mathcal{U}_i[t] \quad i = 1, 2, \cdots, m \tag{5}$$

とした．本稿では DM_i に対して $U_j[t]\,(j \neq i)$ を補助情報，上記の写像を符号化関数と呼ぶ．

3 決定理論に基づいたエージェントの決定

3.1 決定関数

本節では，前節で定義した分散協調モデルにおける各エージェントの意思決定を決定理論に基づいて定式化する．各エージェント DM_i は，それぞれ得た情報 $x_i[1,t]$, $\overline{\mathbf{u}}_i[t]$ から，目標に応じた決定をしていく．本稿では，分散協調システム全体の決定関数を以下のように表記する．

$$D\left(x_1^m[1,t], u_1^m[t]\right) = \left\{D_i^t\left(x_i[1,t], \overline{\mathbf{u}}_i[t]\right) \mid i = 1, \cdots, m\right\}. \tag{6}$$

ここで，時刻 t におけるエージェント DM_i の決定関数を $D_i^t\left(x_i[1,t], \overline{\mathbf{u}}_i[t]\right)$ とした．

エージェントの決定関数によって，さまざまな問題を考えることができる．例えば，全エージェントに対して未知である $x_{m+1}^n[1,t]$ の分布関数，あるいは実現値を推定する問題や，符号化された補助情報 $\overline{\mathbf{u}}_i[t]$ の元の系列 $x_i[1,t]$ の分布関数，あるいは実現値を推定する問題などがあり，全エージェントに対して未知である値の推定問題は，確率分布のパラメータ推定，符号化された補助情報の元の系列を推定する問題は，情報源符号化における復号としてみることができる．

3.2 期待損失関数

分散協調システムの出力に対する損失を，各エージェントが決定した結果に対し，ある距離を導入して評価する．決定の損失として良く用いられる距離としては，0-1 損失，二乗誤差損失，対数損失などがある [2]．ここで，式 (6) の決定関数に対する損失関数を以下のように表記する．

$$L^t\left(D\left(x_1^m[1,t], u_1^m[t]\right), x_1^n[1,t]\right). \tag{7}$$

また，期待損失関数は以下のように定義される．

$$EL^t\left(D\left(X_1^m[1,t], U_1^m[t]\right), X_1^n[1,t]\right) = \sum_{\mathcal{X}_1^t \cdots \mathcal{X}_n^t} \sum_{\mathcal{U}_1[t] \cdots \mathcal{U}_m[t]} P\left(x_1^n[1,t], u_1^m[t]\right) \times L\left(D\left(x_1^m[1,t], u_1^m[t]\right), x_1^n[1,t]\right), \tag{8}$$

$$P\left(x_1^n[1,t], u_1^m[t]\right) = \Pr\left\{X_1^n[1,t] = x_1^n[1,t], U_1^m[t] = u_1^m[t]\right\}. \tag{9}$$

ここで，符号化関数 $\varphi_i^t\,(i = 1, \cdots, m)$ が与えられると，この期待損失関数を最小化する決定関数を求めることができる．

また，情報交換のコスト，すなわち符号化関数 φ_i^t に対するコストを以下のように表記する．

$$C_i^t\left(x_i[1,t], u_i[t]\right), \quad i = 1, \cdots, m.$$

期待コストは以下のように定義される．

$$EC_i^t\left(X_i[1,t], U_i[t]\right) = \sum_{\mathcal{X}_i^t} \sum_{\mathcal{U}_i[t]} P\left(x_i[1,t], u_i[t]\right) \times C_i^t\left(x_i[1,t], \varphi_i^t\left(x_i[1,t]\right)\right). \tag{10}$$

ここで定義した式 (8) と式 (10) との間には有歪み情報源符号化におけるレートと歪みの関係と同様にトレードオフの関係があると予想される．

次節では，これまでに定義した分散協調問題において，$n = 3$, $m = 2$ の場合，すなわちエージェントが 2 人の場合に具体的な損失関数や情報交換のコストを与え，そのときの最適な意思決定，および情報交換について考える．

4 分散協調問題の例

4.1 最適な決定と期待損失関数

例 4.1 各エージェントが共に観測できない $x_3[1,t]$ の条件付き確率分布 $P\left(x_3[1,t] \mid x_1^2[1,t]\right)$ を推測する問題を考え，損失関数を以下のような対数損失として定義する．

$$L^t\left(D^t\left(x_1^2[1,t], u_1^2[t]\right), x_1^3[1,t]\right) = a \log \frac{P\left(x_3[1,t] \mid x_1^2[1,t]\right)}{D_1^t\left(x_1[1,t], \varphi_2^t\left(x_2[1,t]\right)\right)} + b \log \frac{P\left(x_3[1,t] \mid x_1^2[1,t]\right)}{D_2^t\left(x_2[1,t], \varphi_1^t\left(x_1[1,t]\right)\right)}. \tag{11}$$

ここで，a, b は非負の実数．この損失関数に対する期待損失関数を最小にする決定は以下で与えられる．

定理 4.1 任意の符号化関数 φ_1^t, φ_2^t に対し，式 (11) の期待値を最小にする決定関数は以下のような事後確率で与えられる．

$$D^{t(*)}\left(x_1^2[1,t], u_1^2[t]\right) = \left\{D_1^{t(*)}\left(x_1[1,t], u_2[t]\right), D_2^{t(*)}\left(x_2[1,t], u_1[t]\right)\right\}, \tag{12}$$

$$\begin{aligned} D_1^{t(*)}\left(x_1[1,t], u_2[t]\right) &= \sum_{x_2[1,t] \in \mathcal{X}_2^t} P\left(x_3[1,t] \mid x_1^2[1,t]\right) \times P\left(x_2[1,t] \mid x_1[1,t], u_2[t]\right) \\ &= P\left(x_3[1,t] \mid x_1[1,t], u_2[t]\right) \\ &= P\left(x_3[1,t] \mid x_1[1,t], \varphi_2^t\left(x_2[1,t]\right)\right), \end{aligned} \tag{13}$$

$$\begin{aligned} D_2^{t(*)}\left(x_2[1,t], u_1[t]\right) &= \sum_{x_1[1,t] \in \mathcal{X}_1^t} P\left(x_3[1,t] \mid x_1^2[1,t]\right) \times P\left(x_1[1,t] \mid x_2[1,t], u_1[t]\right) \\ &= P\left(x_3[1,t] \mid x_1[1,t], u_1[t]\right) \\ &= P\left(x_3[1,t] \mid x_1[1,t], \varphi_1^t\left(x_1[1,t]\right)\right). \end{aligned} \tag{14}$$

□

符号化関数 φ_1^t, φ_2^t が与えられている場合，定理 4.1 の決定関数に関する期待損失関数は，

$$\begin{aligned} &EL^t\left(D^{t(*)}\left(X_1^2[1,t],U_1^2[t]\right),X_1^3\right) \\ &\quad = aI\left(X_3[1,t];X_2[1,t]\mid X_1[1,t]\right) \\ &\qquad -aI\left(X_3[1,t];U_2[t]\mid X_1[1,t]\right) \\ &\qquad +bI\left(X_3[1,t];X_1[1,t]\mid X_2[1,t]\right) \\ &\qquad -bI\left(X_3[1,t];U_1[t]\mid X_2[1,t]\right) \end{aligned} \tag{15}$$

となり，このような分散協調問題において符号化関数が与えられた場合，各エージェントが最適な決定をしたときのベイズリスクを 4 つの相互情報量で表現できることが示せた．また，

$$\begin{aligned} 0 &\leq I\left(X_3[1,t];U_2[t]\mid X_1[1,t]\right) \\ &\leq I\left(X_3[1,t];X_2[1,t]\mid X_1[1,t]\right), \end{aligned} \tag{16}$$

$$\begin{aligned} 0 &\leq I\left(X_3[1,t];U_1[t]\mid X_2[1,t]\right) \\ &\leq I\left(X_3[1,t];X_1[1,t]\mid X_2[1,t]\right), \end{aligned} \tag{17}$$

が任意の符号化関数 φ_1^t,φ_2^t に対して成り立っているので，期待損失関数は，

$$\begin{aligned} 0 &\leq EL^t\left(D^{t(*)}\left(X_1^2[1,t],U_1^2[t]\right),X_1^3[1,t]\right) \\ &\leq aI\left(X_3[1,t];X_2[1,t]\mid X_1[1,t]\right) \\ &\qquad +bI\left(X_3[1,t];X_1[1,t]\mid X_2[1,t]\right) \end{aligned} \tag{18}$$

の範囲に値をとることがわかる．

以上より，エージェント DM_1 が最適な情報交換をエージェント DM_2 へ行うためには，$X_2[1,t]$ の条件付きで $X_3[1,t]$ と $X_1[1,t]$ の相関が相互情報量の意味で高くなるように φ_1^t を構成すればよいことがわかる． □

例 4.2 エージェント DM_1, DM_2 が，それぞれ $x_2[1,t]$, $x_1[1,t]$ を推測する問題を考え，損失関数を以下のような二乗誤差損失として定義する．

$$\begin{aligned} &L^t\left(D^t\left(x_1^2[1,t],u_1^2[t]\right),x_1^2[1,t]\right) \\ &\quad = a\left(D_1^t\left(x_1[1,t],u_2[t]\right)-x_2[1,t]\right)^2 \\ &\qquad +b\left(D_2^t\left(x_2[1,t],u_1[t]\right)-x_1[1,t]\right)^2. \end{aligned} \tag{19}$$

ここで，a,b は非負の実数であり，この損失関数に対する期待損失関数を最小にする決定関数は以下の定理で与えられる．

定理 4.2 任意の符号化関数 φ_1^t, φ_2^t に対し，式 (19) の期待値を最小にする決定関数は以下のような事後期待値で与えられる．

$$\begin{aligned} &D^{**}\left(x_1^2[1,t],u_1^2[t]\right) \\ &\quad = \left\{D_1^{t(**)}\left(x_1[1,t],u_2[t]\right),D_2^{t(**)}\left(x_2[1,t],u_1[t]\right)\right\}, \end{aligned} \tag{20}$$

$$\begin{aligned} &D_1^{t(**)}\left(x_1[1,t],u_2[1,t]\right) \\ &\quad = \sum_{\mathcal{X}_2^t} x_2[1,2]P\left(x_2[1,t]\mid x_1[1,t],u_2[t]\right), \end{aligned} \tag{21}$$

$$\begin{aligned} &D_2^{t(**)}\left(x_2[1,t],u_1[t]\right) \\ &\quad = \sum_{\mathcal{X}_1^t} x_1[1,t]P\left(x_1[1,t]\mid x_2[1,t],u_1[t]\right). \end{aligned} \tag{22}$$

□

□

注意 4.1 例 4.1 は $x_3[1,t]$ の条件付き確率分布を推測する問題であったが，$x_3[1,t]$ 自体を推測する問題も考えることができ，この場合，例 4.2 と同様に $x_3[1,t]$ の事後期待値が最適な決定となる．また，例 4.2 は直接 $x_1[1,t]$, $x_2[1,t]$ を推測する問題であるが，分布を推測する問題で，損失関数が対数損失の場合は，例 4.1 と同様の結果となる． □

4.2 最適な符号化関数

前節では，符号化関数が与えられているもとで，期待損失関数を最小にする決定について述べた．本節では，定理 4.1 と定理 4.2 の決定関数による期待損失関数を最小にする符号化関数について考える．ここでは，符号化関数に対する期待コストを平均符号長とし，以下のように表記する．

$$EC_i^t\left(X_i[1,t],U_i[t]\right) = ER_i^t\left(U_i[t]\right). \tag{23}$$

ここで，$R_i^t\,(i=1,2)$ は，それぞれ $u_1[t]$, $u_2[t]$ に対する符号長関数とした．

前述のようにエージェントの意思決定に対する損失と情報交換に対するコストの間には，レート歪み関数のようなトレードオフ関係があると考えられるので，式 (23) の期待コストに以下のような制約を与え，そのもとで期待損失関数を最小化することを考える．すなわち，制約

$$\begin{aligned} &ER_1^t\left(U_1[t]\right) \leq R_1^t, \\ &ER_2^t\left(U_2[t]\right) \leq R_2^t \end{aligned} \tag{24}$$

のもと，期待損失関数を最小にするような符号化関数を構成する．また，ここでは符号化関数 φ_1^t, φ_2^t を，以下のような条件付き確率分布

$$Q_1^t\left(u_1[t]\mid x_1[1,t]\right),\ Q_2^t\left(u_2[t]\mid x_2[1,t]\right)$$

とする．

交換情報 $U_i[t]=\varphi_i^t\left(X_i[1,t]\right)$ を，もう一方のエージェントに送るときに必要な平均符号長は $ER_i^t\left(\varphi_i^t\left(X_i[1,t]\right)\right)$ となるが，その最小値はシャノンの符号化定理 [1] より確率変数 U_i のエントロピー $H\left(U_i\right)$ と一致することがわかっている．したがって，符号長関数が，

$$R_1^t\left(u_1[t]\right) = -\log P\left(u_1[t]\right), \tag{25}$$

$$R_2^t\left(u_2[t]\right) = -\log P\left(u_2[t]\right) \tag{26}$$

のとき平均符号長が最小になるので，以下では符号長関数を上式として定義する．ここで，

$$P\left(u_1[t]\right) = \sum_{\mathcal{X}_1^t} Q_1^t\left(u_1[t]\mid x_1[1,t]\right)P\left(x_1[1,t]\right), \tag{27}$$

$$P\left(u_2[t]\right) = \sum_{\mathcal{X}_2^t} Q_2^t\left(u_2[t]\mid x_2[1,t]\right)P\left(x_2[1,t]\right) \tag{28}$$

とした．

式 (11) と式 (19) の損失関数は，各エージェントの決定に関する損失の線形和となっているので，符号化関数 φ_1^t, φ_2^t は，それぞれ独立に構成することができる．よって，例 4.1 における最適な符号化関数は，

$$H(U_1) \leq R_1,\ H(U_2) \leq R_2 \tag{29}$$

を満たし，

$$\begin{aligned}
&EL\left(D^{t(*)}\left(X_1^2[1,t], U_1^2[t]\right), X_1^3[1,t]\right) \\
&= \sum_{\mathcal{X}_1^t \cdots \mathcal{X}_3^t} \sum_{\mathcal{U}_1[t], \mathcal{U}_2[t]} P\left(x_1^3[1,t], u_1^2[t]\right) \\
&\qquad \times L\left(D^{t(*)}\left(x_1^2[1,t], u_1^2[t]\right), x_1^3[1,t]\right) \\
&= \sum_{\mathcal{X}_1^t \cdots \mathcal{X}_3^t} \sum_{\mathcal{U}_1[t], \mathcal{U}_2[t]} Q_1^t\left(u_1[t] \mid x_1[1,t]\right) Q_2^t\left(u_2[t] \mid x_2[1,t]\right) \\
&\qquad \times P\left(x_1^3[1,t]\right) L\left(D^{t(*)}\left(x_1^2[1,t], u_1^2[t]\right), x_1^3[1,t]\right) \\
&= aI\left(X_3[1,t]; X_2[1,t] \mid X_1[1,t]\right) \\
&\qquad -aI\left(X_3[1,t]; U_2[t] \mid X_1[1,t]\right) \\
&\quad +bI\left(X_3[1,t]; X_1[1,t] \mid X_2[1,t]\right) \\
&\qquad -bI\left(X_3[1,t]; U_1[t] \mid X_2[1,t]\right)
\end{aligned} \tag{30}$$

を最小にする $Q_1^t\left(u_1[t] \mid x_1[1,t]\right)$，$Q_2^t\left(u_2[t] \mid x_2[1,t]\right)$ を構成することで与えられる．例 4.2 に関しても同様である．

また，R_1^t，および R_2^t の最大値は期待損失が 0 になる符号化関数を構成することで求めることができる．式 (29) のような期待コストの制約の場合，R_1^t，R_2^t の最大値は，簡単な計算から，それぞれ $H\left(X_1[1,t]\right)$ と $H\left(X_2[1,t]\right)$ で与えられる．このことから，最小期待損失の最大値は (R_1, R_2) − 平面の座標 $(0,0)$ で最大となり，座標

$$\left(H\left(X_1[1,t]\right), H\left(X_2[1,t]\right)\right)$$

より大きい点では最小の 0 になることがわかる．

5 数値計算

図 2　数値計算結果 $(a = b = 1, t = 1)$

エージェントの意思決定による損失と情報交換によるコストのトレードオフ関係をみるため，式 (29) の制約のもと例 4.1 における式 (30) の期待損失関数の最小化を非線形最適化手法を用いて数値計算した．時刻 1 における結果の一部を図 2 に示す．

また，条件は以下の通りで，有限集合を

$$\mathcal{X}_i = \{0,1\},\quad i = 1,2,3, \tag{31}$$

$$\mathcal{U}_j[1] = \{0,1\},\quad j = 1,2, \tag{32}$$

情報源の確率分布 $P\left(x_1[1], x_2[1], x_3[1]\right)$ を

$$\begin{aligned}
P(0,0,0) &= \frac{1}{16},\ P(0,0,1) = \frac{1}{8}, \\
P(0,1,0) &= \frac{1}{16},\ P(0,1,1) = \frac{1}{4}, \\
P(1,0,0) &= \frac{1}{16},\ P(1,0,1) = \frac{1}{4}, \\
P(1,1,0) &= \frac{1}{8},\ P(1,1,1) = \frac{1}{16}.
\end{aligned}$$

として計算を行った．

数値計算により，R_1^t，R_2^t，および最小期待損失の最大値が前述したものと一致していることが確かめられた．さらに，この例の場合の期待損失関数は，R_1^t，R_2^t に関して上に凸な関数になっている．また，例 4.2 のように二乗誤差損失を決定の損失関数として与えると，レート歪み関数と同様，下に凸な関数になる．

6 まとめ

本稿では，多端子情報源符号化と統計的決定理論に基づいて分散協調問題をモデル化し，各エージェントの意思決定を定式化した．また，対数損失と二乗誤差損失による損失関数に対して，最適な決定関数を導出し，そのときの対数損失に関する期待損失関数が相互情報量で表現できることを示した．さらに，情報交換によるコストと期待損失関数とのトレードオフ関係を数値計算により明らかにした．今度の課題としては具体的な符号化関数の構成，およびモデルの一般化などがあげられる．

7 謝辞

本研究の一部は早稲田大学特定課題研究助成費 (2001A-570)，文部省科学研究費基盤 (C)(No.12650400) の援助による．

参考文献

[1] T.M.Cover and J.A.Thomas, "Elements of information theory," John Wiley & Sons, New York, 1991.

[2] J.O.Berger, "statistical decision theory and Bayesian analysis," Springer-Verlag, New York, 1985.

[3] T.H.Han and S.Amari, "Parameter estimation with multiterminal data compression," *IEEE Trans. Inf. Theory*, vol.41, no.6, pp.1802-1833, Nov. 1995.

[4] S.Amari and T.H.Han, "Statistical inference under multiterminal rate restrictions: A differential geometrical approach," *IEEE Trans. Inf. Theory*, vol.35, no.2, pp.217-227, Mar. 1989.

[5] Z. Zhang and T. Beger, "Estimation via compressed information," *IEEE Trans. Inf. Theory*, vol.34, no.2, pp.198-211, Mar. 1988.

An Alternate Algorithm for Calculating Generalized Posterior Probability and Decoding

Toshiyasu Matsushima
Waseda University
Shinjuku, Tokyo, JAPAN
e-mail: toshi@matsu.mgmt.waseda.ac.jp

Tomoko K. Matsushima
Polytechnic University,
Sagamihara, JAPAN

Shigeichi Hirasawa
Waseda University,
Shinjuku, Tokyo, JAPAN

I. Introduction

The accurate and efficient calculation of ordinary and generalized posterior distributions is an important problem in the several research fields such as decoding, AI, statistics and statistical mechanics. The condition of generalized posterior distributions is not given by the deterministic values such as $X = x$, but by the distributions such as $P(X = x) = p_x$. If the condition is $P(X = x) = 1$ then the generalized posterior distribution is an ordinary posterior distribution.

In this paper, a procedure using the sum of the e-projection [1] vectors shall be proposed. Since the procedure is suitable for parallel algorithms, an alternate algorithm for calculating generalized posterior distributions on log linear models is given by the procedure. The proposed algorithm works well for the codes with short loops.

II. A calculation procedure for generalized posterior probability

A generalized posterior distribution $P(X|S_1, \ldots, S_l)$ given the conditions that S_i is $P(y_i) = p_i$ is calculated by iterating the e-projection on the submanifold satisfies the each condition S_i from a prior distribution [4]. In this paper, another procedure for calculating generalized posterior distributions shall be proposed. A generalized posterior distribution $P(X|S_1, \ldots, S_l)$ is deduced from iterating the translation by the sum of the e-projection vectors on the submanifold of S_i from a prior distribution $P(X)$.

III. An alternate algorithm on log linear models

A log linear mode is represented by the following joint probability.

$$P(x_1, \ldots, x_n) = \alpha \frac{q(N_1)q(N_2)\cdots q(N_I)}{q(D_1)^{R_1} \cdots q(D_J)^{R_J}}, \quad (1)$$

where X is a discrete random variable, $q(N_i) = q(x_{1(i)}, \ldots, x_{n_i(i)})$ and $q(D_j) = q(x_{1(j)}, \ldots, x_{d_j(j)})$.

$t(N_i) = \{x_{1(i)}, \ldots, x_{n_i(i)}\}\ i \in \{1, \ldots, I\}$ and $t(D_j) = \{x_{1(j)}, \ldots, x_{d_j(j)}\}\ j \in \{1, \ldots, J\}$ are called numerator term and denominator term respectively. Abbreviate $t(N_i), t(D_j)$ to N_i, D_j respectively. In Formula(1), $N_i \subseteq N_k$ is not satisfied for any $i \in \{1, \ldots, I\}$ and any $k \neq i$.

The neighboring node set of a denominator term node is defined as follows: If for any l, $D_j \subset D_l$ is not satisfied then $S^N(D_j) = \{N_i | D_j \subset N_i\}$, else $S^N(D_j) = \{N_i | (D_j \subset N_i) \wedge (\forall N_k \in S^N(D_j)(k \neq i) N_i \cap N_k \in S^S(D_j))\}$, where $S^S(D_j) = \{D_l | D_j \subset D_l\}$. The neighboring node set of a numerator term node N_i is defined by $S^D(N_i) = \{D_j | N_i \in S^N(D_j)\}$.

The above mentioned models may be represented by a graph using two types of nodes that is given by connecting each denominator term node to all numerator term nodes in its neighboring node set with arcs. The denominator term nodes differ from intersection nodes in Junction graphs[1][2]. Although an intersection node is connected to two clique nodes, a denominator term node may be connected to more than two numerator term nodes.

We shall propose a new alternate propagation algorithm on the graphs for calculating marginal posterior probabilities.

[1] Let M_S be a m-flat submanifold in a distribution space M, and let a distribution p be a point in M. Iff the point q is given by the e-geodesic projection of p to M_S, $q = \arg\min_{q \in M} D(p\|q)$ where $D(p\|q)$ is the Kullback information between p and q.

Two contrary directed propagations repeat alternately in the algorithm. In the first type of message propagation, all numerator term nodes propagate each message to the all connected denominator term nodes. In the contrary message propagation, all denominator term nodes propagate each message to all connected numerator term nodes.

$$q_{t+1}(D_j) = \frac{\prod_{N_i \in S^N(D_j)} \sum_{x \notin D_j} q_t(N_i)}{q_t(D_j)^{R_j}}. \quad (2)$$

$$q_{t+1}(N_i) = \prod_{D_j \in S^D(N_i)} \frac{q_{t+1}(D_j)}{\sum_{x \notin D_j} q_t(N_i)} q_t(N_i). \quad (3)$$

The procedure may be calculated by the one formula that is compounded from the two formulas mentioned above. The compounded procedure is similar to Generalized Distribution Low(GDL)[1]in the case of the calculation for ordinary posterior probability.

The marginal posterior probability of each random variable is calculated by

$$q_t(x) = \sum_{xf \neq x} \frac{\prod_{\{N_i | x \in N_i\}} q_t(N_i)}{\prod_{\{D_j | x \in D_j\}} q_t(D_j))^{R_j}}. \quad (4)$$

If the graph has no loops, the propose algorithm halts and calculates exact marginal posterior probabilities.

The algorithm is regarded as an alternate type algorithm for HUGIN algorithm[2]. Although HUGIN algorithm only works on Junction trees, the proposed algorithm can work on the Junction graphs. And the probability model class for which the proposed algorithm calculates exact marginal posterior probability is wider than that of the sum-products algorithm[3]. Needless to say, HUGIN algorithm and the sum-products algorithm can not calculate generalized posterior distributions. From these view points, the proposed algorithm is regarded as a generalization of these algorithms.

IV. An application to decoding

Although a main problem of decoding does not need the calculation of generalized posterior probability, the proposed algorithm may be applied to decoding. The previous research papers reported the sum-products algorithm does not work well for the factor graph with short loops. In the case that short loops overlap together, the clustering on factor graph raises up a computational explosion.

However, the proposed algorithm can guarantee the exact marginal posterior probability with respect to the probability model that has overlapped short loops of length 4 in a factor graph. For example, the proposed algorithm is applied to the LDPC codes of short codelength $n = 500$ with short loops that is constructed randomly. It is shown from several simulations that the bit and the block error rate of the proposed algorithm are better than those of the sum-products algorithm. The proposed algorithm also works well for tail biting codes.

References

[1] S.M. Aji, R.J. Mcliece, *The Generalized Distributive Law*, IEEE Trans. IT, Vol.46 No.2, 2000.

[2] F. V. Jensen, *An introduction to Baycsian networks*, University College London Press, London, 1996.

[3] F.R. Kschischang, B.J. Fey and H. Loeliger, *Factor Graphs and the Sum-Product Algorithm*, IEEE Trans. IT, Vol.47 No.2, 2001.

[4] T. Matsushima, T.K. Matsushima and S. Hirasawa *An Iterative Algorithm for Calculating Posterior Probability and Model Representation* , Proceedings of IEEE Int. Symp. on IT, 2001.

The 25th Symposium on Information Theory
and Its Applications (SITA2002)
Ikaho, Gunma, Japan, Dec. 10–13, 2002

Calculation of Generalized Posterior Distribution on Junction Graphs

Toshiyasu Matsushima* Tomoko K. Matsushima† Shigeichi Hirasawa*

Abstract—Although there are serial and alternate algorithms for calculating ordinary posterior distributions, there are only serial algorithms for generalized posterior distributions. First, we propose parallel iterative procedures to calculate joint generalized posterior probability. Secondly, we propose efficient alternate propagation algorithms based on the procedures mentioned above for calculating marginal generalized posterior distributions on extended junction graphs (EJG).

Keywords—posterior probability, probabilistic reasoning, iterative decoding, graphical model

1 Introduction

The posterior probability $P(Y|X = x)$ given the evidence $X = x$ is calculated by the ordinary probabilistic reasoning such as Belief Propagation (BP) [8] and HUGIN algorithm[4]. In this case, the evidence is represented by the deterministic value of some random variable such as $X = x$. In the case that an evidence is given by the distribution of random variables such as $P(X = x) = p_x$, this type of evidence is called distribution-evidence or soft-evidence in the previous research[9].

The probabilistic reasoning given distribution-evidence is formalized as generalized probabilistic reasoning[6]. The research shows that the generalized probabilistic reasoning is the same problem as minimizing K-L information to a prior distribution under the condition that some marginal distributions are given. The results calculated by the generalized probabilistic reasoning are also interpreted as the generalized posterior distributions given some marginal distributions, because the values play the same role as posterior distributions do in statistical inference.

We cannot calculate even joint generalized posterior distributions as easily as ordinary posterior probability. Iterative Proportional Fitting Procedure (IPFP) or Iterative Scaling Procedure (ISP) [2][3]is an iterative algorithm to calculate generalized posterior distribution. Efficient algorithms based on IPFP have been proposed for calculating marginal generalized posterior distributions on decomposable probabilistic models[6]. Since IFPF is a serial procedure, these algorithms are serial algorithms.

Although there are serial and alternate algorithms for ordinary posterior distributions[1] , there are only serial algorithms for generalized posterior distributions. Therefore, first, we propose parallel iterative procedures to calculate joint generalized posterior probability. We can consider two methods to represent a target generalized posterior probability in the procedure. One is the method that the generalized posterior probability is directly represented by a joint probability. The other is the method that the target generalized posterior probability is represented by the prior joint probability and marginal parameters.

Secondly, we propose efficient alternate propagation algorithms based on the proposed procedures for calculating marginal generalized posterior distributions on extended junction graphs (EJG). The intersection nodes in an EJG differ from the intersection nodes in a junction graph (JG)[1][4]. Although an intersection node in a JG is connected to two clique nodes, an intersection node in an EJG may be connected to more than two clique nodes. Two contrary directed propagations repeat alternately in the proposed algorithm. If the graph has no loops, the proposed algorithm halts and calculates exact marginal generalized posterior probabilities.

2 Generalized posterior probability and IPFP

First, we formalize a generalized probabilistic reasoning given multi distribution-evidence. Let X_i $i \in I = \{1, \ldots, n\}$ and E_j $j \in I_C \subset I$ be discrete random variables and E_j is called evidence.

Assumption 1 [2] *Each piece of evidence E_j and every X_i $i \in I - \{j\}$ are conditional independent given X_j as follows:*

$$P(X_1, \ldots, X_n, E_j) = \frac{P(X_1, \ldots, X_n)P(X_j, E_j)}{P(X_j)}. \quad (1)$$

Definition 1 *Generalized probabilistic reasoning is defined as the calculation of the probability $P(X_{k_1}, \ldots, X_{k_{|I_{out}|}}|E^{I_C} = e^{I_C})$, $k_i \in I_{out} \subset I$ from a prior distribution $P(X_1, \ldots, X_n)$ under the condition that evidence $E^{I_C} = e^{I_C}$ is given, where $E^{I_C} = e^{I_C}$ denotes $(E_{j_1} = e_{j_1}, \ldots, E_{j_{|I_C|}} = e_{j_{|I_C|}})$ and a piece of evidence e_j gives the information that the marginal distribution of X_j is $P^*(X_j)$. The information of the evidence e_j is rewritten as $P^*(X_j) = P(X_j|E_j = e_j)$.*

If the distribution $P(X_j|E_j = e_j)$ is the point mass in $X_i = x_i$, i.e., $P(X_j = x_i|E_j = e_j) = 1$, the information from the evidence e_j is the same as "x_i occurred", i.e., $X_i = x_i$. In this case, the defined generalized probabilistic reasoning is identical with the calculation of the probability $P(X_{k_1}, \ldots, X_{k_{|I_{out}|}}|X^{I_C} = x^{I_C})$, where

* Waseda University, Tokyo, JAPAN. Email: toshi@matsu.mgmt.waseda.ac.jp
† Polytechnic University, Sagamihara, JAPAN.

[1] BP and HUGIN algorithm are typical serial algorithms. The sum-product algorithm[5] is a well-known alternate algorithm.

[2] This assumption is identical with the assumption of the research[9].

$X^{I_C} = x^{I_C}$ denotes $(X_{j_1} = x_{j_1}, \ldots, X_{j_{|I_C|}} = x_{j_{|I_C|}})$. It is the same as the ordinary probabilistic reasoning. Thus, the defined generalized probabilistic reasoning includes the ordinary probabilistic reasoning in the special case.

The distributions calculated by the generalized probabilistic reasoning is also interpreted as generalized posterior distributions given marginal distributions $P^*(X_j)$ instead of given strict values $X^{I_C} = x^{I_C}$.

Theorem 1 *[2][6] Let M_C be the set of distributions on the random variables $(X_1, \ldots, X_n)$ that satisfies the marginal condition $P^*(X_j) = P(X_j|E_j = e_j)$ $j \in I_C$. The posterior distribution $P_{po} = P(X_1, \ldots, X_n|E^{I_C} = e^{I_C})$ that is induced by the generalized probabilistic reasoning defined by Definition 2.1 under Assumption 2.1 is given by*

$$P_{po} = \arg \min_{P \in M_C} I(P||P_{pr}), \tag{2}$$

where P_{pr} is a prior distribution $P(X_1, \ldots, X_n)$.

IPFP can be applied to generalized probabilistic reasoning, because IPFP calculates the distribution that is closest to a prior distribution with K-L information under the restriction of marginal distribution. We show the IFPF of generalized probabilistic reasoning for calculating $P(X_1, \ldots, X_n|E^{I_C} = e^{I_C})$. It corresponds to the case $I_{out} = I$ in Definition 2.1.

[Procedure 1: IFPF]
begin
$P(X_1, \ldots, X_n) := P_{pr}(X_1, \ldots, X_n)$;
while $\exists_{j \in I_C} P(X_j) \neq P^*(X_j)$ do
  begin
  Pick up X_j from $\{X_j | P(X_j) \neq P^*(X_j), j \in I_C\}$;
  $P(X_1, \ldots, X_n) := P(X_1, \ldots, X_n) \frac{P^*(X_j)}{P(X_j)}$;
  end
$P_{po}(X_1, \ldots, X_n) := P(X_1, \ldots, X_n)$;
end

Lemma 1 *Procedure 1 halts. Then the value calculated by Procedure 1 converges to $P(X_1, \ldots, X_n|E^{I_C} = e^{I_C})$.*

3 A parallel procedure for calculating generalized posterior probability

Procedure 1 is a serial iterative algorithm for calculating generalized posterior probability. Under the same condition of Procedure 1, a parallel procedure is proposed as follows: Although Procedure 1 renews a distribution by adjusting its marginal to one marginal restriction at each cycle, this procedure adjusts it to all marginal restrictions simultaneously.

[Procedure 2A]
begin
$P(X_1, \ldots, X_n) := P_{pr}(X_1, \ldots, X_n)$;
while $\exists_{j \in I_C} P(X_j) \neq P^*(X_j)$ do
  begin
  $P(X_1, \ldots, X_n) := P(X_1, \ldots, X_n) \prod_{j \in I_C} \frac{P^*(X_j)}{P(X_j)}$;
  end
$P_{po}(X_1, \ldots, X_n) := P(X_1, \ldots, X_n)$;
end

Although we keep the information of the target distribution as a joint probability $P(X_1, \ldots, X_n)$ in the procedure, we can keep it with the marginal parameter $\beta(X_j), j \in I_C$, because the target distribution can be represented by using the marginal parameter and the prior distribution[3]. We can rewrite Procedure 2A by using the marginal parameter representation as follows:

[Procedure 2B]
begin
$\beta(X_j) := 0, j \in I_C$;
while $\exists_{j \in I_C} P(X_j) \neq P^*(X_j)$ do
  begin
  $\gamma(X_j) := \sum_{i \neq j} P_{pr}(X_1, \ldots, X_n) \prod_{j \in I_C} \beta(X_j), j \in I_C$;
  $\beta(X_j) := \frac{P^*(X_j)}{\gamma(X_j)}, j \in I_C$;
  end
$P_{po}(X_1, \ldots, X_n) := P_{pr}(X_1, \ldots, X_n) \prod_{j \in I_C} \beta(X_j)$;
end

Theorem 2 *Both Procedure 2A and 2B converge to the distribution that Procedure 1 converges to.*

4 Efficient algorithms on extended junction graphs

4.1 Representation of probabilistic model and extended junction graphs

Efficient alternate propagation algorithms based on Procedure 2A and 2B are proposed for calculating marginal generalized posterior probability on an extended junction graph (EJG) in this section. An EJG is almost the same as an ordinary junction graph (JG), but the intersection nodes differ. Although an intersection node in a JG is connected to two clique nodes, an intersection node in an EJG may be connected to more than two clique nodes. If all intersection nodes are eliminated from a JG, the remaining graph is also a normal graph[3]. However, if all intersection nodes are eliminated from an EJG, the remaining graph might be a hyper graph.

An EJG is defined by a clique node set $S_N = \{N_1, N_2, \ldots N_{n_N}\}$, an intersection node set $S_D = \{D_1, \ldots, D_{n_D}\}$ and the neighboring node set $S^N(D_m)$ of every intersection node $D_m, m = 1, \ldots, n_D$. Each intersection node is connected to all clique nodes in its neighboring node set with arcs in an EGJ.

An EJG and a JG are applied to the representation of the probability model whose joint distribution factors into a product of several local functions of some subset of random variables. Two typical types of joint

[3] In the paper[1], junction trees(JT) are represented by eliminating all intersection nodes.

distribution represented by EJGs and JGs are shown as follows.

The first type is given by

$$P(X_1,\ldots,X_n) = \frac{P(N_1)P(N_2)\cdots P(N_{n_N})}{P(D_1)^{R_1}\cdots P(D_{N_D})^{R_{N_D}}}, \quad (3)$$

where X_i is a discrete random variable, $P(N_l) = P(X_{i_1(l)},\ldots,X_{i_{n(l)}(l)})$ and $P(D_j) = P(X_{i_1(m)},\ldots,X_{i_{n(m)}(m)})$.

$t(N_l) = \{X_{i_1(l)},\ldots,X_{i_{n(l)}(l)}\}$, $l \in \{1,\ldots,n_N\}$ and $t(D_m) = \{X_{i_1(m)},\ldots,X_{i_{n(m)}(m)}\}$, $m \in \{1,\ldots,n_D\}$ are called clique elements and intersection elements respectively. Abbreviate $t(N_l)$, $t(D_m)$ to N_l, D_m respectively. In Formula(1), $N_l \subseteq N_{l'}$ is not satisfied for any $l \in \{1,\ldots,n_N\}$ and any $l' \neq l$.

The distributions that can be represented by BN are included in this of distributions. The alternate algorithm based on Procedure 2A is applied to this of distributions.

The other typical type of joint distribution is represented by

$$P(x_1,\ldots,x_n) = \alpha q(N_1)q(N_2)\cdots q(N_{n_N}). \quad (4)$$

The alternate algorithm based on Procedure 2B is applied to this of distributions.

If a neighboring node set $S^N(D_m)$ of every intersection node $D_m, m = 1,\ldots,n_D$ is defined in a distribution of the first type, the EJG representing the distribution is uniquely defined. If an intersection node set $S_D = \{D_1,\ldots,D_{n_D}\}$ and a neighboring node set $S^N(D_m)$ of every intersection node $D_m, m = 1,\ldots,n_D$ is defined in a distribution of the second type, an EJG representing the distribution is uniquely defined,

There are several methods to construct JGs or JTs from a joint distribution of these types. An EJG can also be constructed from the joint probability by the similar methods to the previous methods for JGs or JTs.

4.2 Alternate algorithm on EJG

We propose a new alternate propagation algorithms on EJGs for calculating marginal posterior probabilities: $P_{po}(N_l) = \sum_{X \notin N_l} P_{po}(X_1,\cdots,X_n), l = 1,\ldots,n_N$. Two contrary directed propagations repeat alternately in the proposed algorithm. In the first type of message propagation, every intersection node propagates each message to all clique nodes connected with it. In the alternate message propagation, every clique node propagates each message to all intersection nodes connected with it.

First, restricted intersection nodes (r.i.n.) are defined. If the element of an intersection node is equivalent to a restricted random variable as $D_m = \{X_j\}, j \in I_C$, the intersection node is called a restricted intersection node. If there does not exist an intersection node satisfying $D_m = \{X_j\}, j \in I_C$, the restricted intersection node corresponding to every such restricted random variable X_j is produced and connected to an arbitrary clique node N_l satisfying $X_j \in N_l$.

[Algorithm 1A]: Alternate message propagation (joint probability representation)

The alternate propagation algorithm, using the joint probability representation for the first type of distribution, is proposed by applying Procedure 2A. The values of $P_t(D_m)$ and the values of $P_t(N_l)$ are held in every intersection node and every clique node respectively at each step t.

The message from each intersection node D_m is calculated by the following formula. If the random variable of the intersection node D_m is restricted by its marginal distribution, the message is the same as the restricted marginal.

$$P_{t+1}(D_m) = \begin{cases} P^*(D_m) & \text{if } D_m \text{ is a r.i.n.} \\ \frac{\prod_{N_l \in S^N(D_m)} P_t^{N_l}(D_m)}{P_t(D_m)^{R_m}} & \text{otherwise .} \end{cases} \quad (5)$$

The message from a clique node N_l to an intersection node D_m is calculated by $P_{t+1}^{N_l}(D_m)$.

$$P_{t+1}(N_l) = \prod_{D_m \in S^D(N_l)} \frac{P_{t+1}(D_m)}{\sum_{x \notin D_m} P_t(N_l)} P_t(N_l). \quad (6)$$

$$P_{t+1}^{N_l}(D_m) = \sum_{X \notin D_m} P_{t+1}(N_l). \quad (7)$$

The marginal posterior probability of each random variable is calculated by

$$P_t(X) = \sum_{Xf \neq X} \frac{\prod_{\{N_l | X \in N_l\}} P_t(N_l)}{\prod_{\{D_m | X \in D_m\}} P_t(D_m))^{R_j}}. \quad (8)$$

[Algorithm 1B]: Alternate message propagation (marginal parameter representation)

The alternate propagation algorithm using the marginal parameter representation is proposed by applying Procedure 2B. The message from an intersection node D_m to a clique node N_l is calculated by

$$q_{t+1}^{N_l}(D_m) = \begin{cases} \frac{P^*(D_m)}{\gamma_t^{N_l}(D_m)} & \text{if } D_m \text{ is a r.i.n.} \\ \prod_{\{N_k \in S^N(D_m) | k \neq l\}} \gamma_t^{N_k}(D_m) & \text{otherwise .} \end{cases} \quad (9)$$

The message from a clique node N_l to an intersection node D_m is calculated by

$$\gamma_t^{N_l}(D_m) = \sum_{x \notin D_m} q(N_l) \prod_{\{D_h \in S^D(N_l) | h \neq m\}} q_t^{N_l}(D_h). \quad (10)$$

The information of $\beta_t(X_j)$ is given by follows. However, the information does not need to calculate marginal posterior distributions.

$$\beta_t(D_m) = \frac{P^*(D_m)}{\prod_{N_l \in S^N(D_m)} \gamma_t^{N_l}(D_m)}. \quad (11)$$

The marginal posterior probability of each random variable is calculated by

$$P_{t+1}(N_l) = q(N_l) \prod_{D_h \in S^D(N_l)} q_t^{N_l}(D_h). \quad (12)$$

[Algorithm 2]: Unique type of message

Two types of message propagation repeat alternately in Algorithm 1A and 1B. The first type is from intersection nodes to clique nodes connected with them. The other type is from clique nodes to intersection nodes connected with them. We can combine these two types of message into one, which propagates from every clique node to all clique nodes connected with it by way of an arbitrary intersection node. The procedure exchanging a unique type of message between connected clique nodes is induced by compounding the two formulas in the alternate algorithm using marginal parameter representation into one formula.

We attach some dummy clique nodes to an original EJG before the procedure is used. A dummy clique node N_i corresponding to every restricted intersection node $D_m = \{X_j\}, j \in I_C$ is produced and connected to the restricted intersection node. The dummy clique node satisfies $N_i = D_m = \{X_j\}$.

The message from a clique node N_i to a clique node N_l connected with it by way of an intersection node D_m is calculated by

$$\mu_{t+1}^{i \to l} = \begin{cases} \frac{P^*(N_i)}{\mu_t^{l \to i}} & \text{if } D_m \text{ is a r.i.n.} \\ \sum_{x \notin D_m} q(N_i) \prod_{N_k \in S_l^N(N_i)} \mu_t^{k \to i} & \text{otherwise,} \end{cases} \quad (13)$$

where $S_l^N(N_i) = \{N_k \in S^N(D_h) | D_h \in S^D(N_i), D_h \notin S^D(N_l), k \neq i\}$.

The marginal posterior probability of each random variable is calculated by

$$P_t(N_l) = q(N_l) \prod_{N_k \in S^N(N_l)} \mu_t^{k \to l} \quad (14)$$

where $S^N(N_l) = \{N_k \in S^N(D_h) | D_h \in S^D(N_l), k \neq l\}$.

Theorem 3 *If an EJG has no loops, Algorithms 1A, 1B and 2 halt and calculate the exact marginal posterior distributions on the EJG.*

5 Consideration and conclusion

We compare the proposed alternate propagation algorithms with previous massage propagation algorithms. First, since generalized posterior distributions include ordinary posterior distributions in a special case, the proposed algorithms can also calculate ordinary posterior distributions.

Secondly, The proposed propagation algorithms can be calculate the exact marginal posterior distributions in the case where the graphs such as BNs, JGs and factor-graphs have no loops, because, the distribution that can be represented by one of these previous graphs with no loops can also represented by a EJG with no loops. From these two points, the proposed algorithms are regarded as a generalization of these previous propagation algorithms.

The proposed algorithm using the joint probability representation is regarded as an alternate type of HUGIN algorithm. Although HUGIN algorithm only works on JTs, the proposed algorithm can work on EJGs. The proposed algorithm of unique message type is regarded as an alternate or full parallel type algorithm of GDL[1]. Alternate propagation algorithms are useful for approximate calculation on graphs with loops.

The proposed algorithms are similar to the sum-product algorithm on the point that both are alternate propagation algorithms. However, the probability model class for which the proposed algorithms calculate exact marginal posterior distributions is wider than that of the sum-product algorithm, because the distribution class represented by JTs is wider than that represented by the factor trees that are constructed by clustering on any factor graphs.

The proposed alternate algorithm is also applied to a decoding problem. From some experimental results, the proposed algorithms work well for the LDPC codes with short loops of length 4 and tail biting codes.

References

[1] S.M. Aji, R.J. Mcliece, *The Generalized Distributive Law*, IEEE Trans. IT, Vol.46 No.2, 2000.

[2] I. Csiszar, *I-divergence geometry of probability distributions and minimization problems*, The Annals of Probability, Vol. 13, No. 1, 146-158, 1975.

[3] C. T. Ireland and S. Kullback, *Contingency tables with given marginals*, Biometrika, Vol. 55, 179-188, 1968.

[4] F.V. Jensen, K.G. Olesen and S.K. Andersen, *An Algebra of Bayesian Belief Universes for Knowledge-Based System*, Networks, Vol.20, 1990.

[5] F.R. Kschischang, B.J. Fey and H. Loeliger, *Factor Graphs and the Sum-Product Algorithm*, IEEE Trans. IT, Vol.47 No.2, 2001.

[6] T. Matsushima, T.K. Matsushima and S. Hirasawa *An Iterative Calculation Algorithm for Posterior Probability*, Proceedings of the 23rd Symposium on Information Theory and Its Applications, 2000.

[7] T. Matsushima, T.K. Matsushima and S. Hirasawa *An Iterative Algorithm for Calculating Posterior Probability and Model Representation* , Proceedings of IEEE Int. Symp. on Information Theory, 2001.

[8] J. Pearl, *Probabilistic reasoning in intelligent systems* Morgan Kaufmann, 1988.

[9] J. Pearl *Jeffry's rule, passage of experiments and Neo-Bayesianism*, Knowledge Representation and Defeasible Reasoning, 245-265, Kluwer Academic Publisher, 1990.